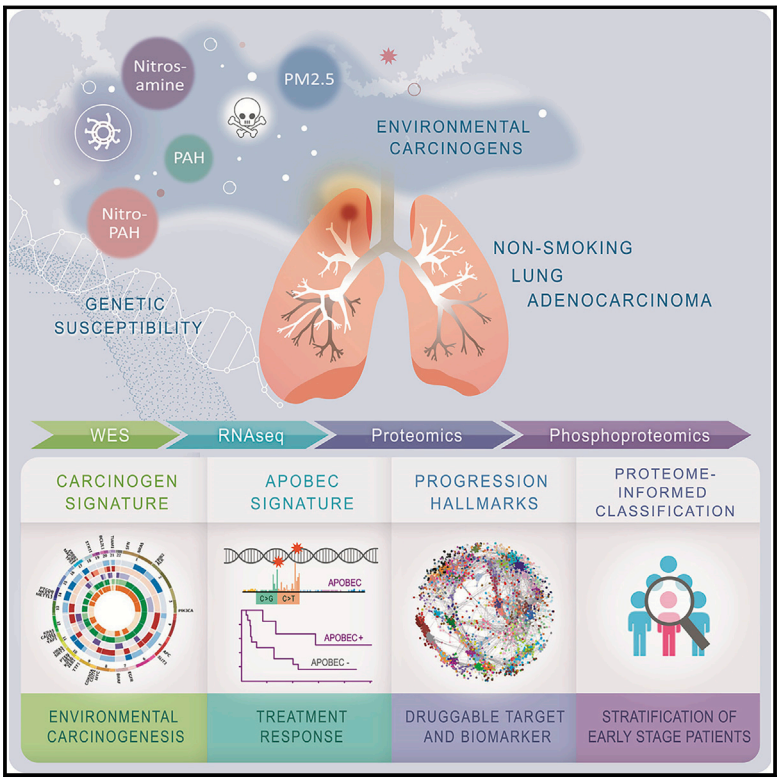
Cell

Proteogenomics of Non-smoking Lung Cancer in East Asia Delineates Molecular Signatures of Pathogenesis and Progression

Graphical Abstract



Highlights

- First deep proteogenomic landscape of non-smoking lung adenocarcinoma in East Asia
- Identified age, sex-related endogenous, and environmental carcinogen mutagenic processes
- Proteome-informed classification distinguished clinical features within early stages
- Protein networks identified tumorigenesis hallmarks, biomarkers, and druggable targets

Authors

Yi-Ju Chen, Theodoros I. Roumeliotis, Ya-Hsuan Chang, ..., Hsuan-Yu Chen, Pan-Chyr Yang, Yu-Ju Chen

Correspondence

was@tmu.edu.tw (C.-L.H.), tsung@iis.sinica.edu.tw (T.-Y.S.), chenjs@ntu.edu.tw (J.-S.C.), slyu@ntu.edu.tw (S.-L.Y.), jyoti.choudhary@icr.ac.uk (J.S.C.), hychen@stat.sinica.edu.tw (H.-Y.C.), pcyang@ntu.edu.tw (P.-C.Y.), yujuchen@gate.sinica.edu.tw (Y.-J.C.)

In Brief

Deep proteogenomic landscape of early stage lung adenocarcinoma in a cohort of mostly non-smokers reveals unique drivers and biomarkers, as well as gender-associated mutagenesis.

Chen et al., 2020, Cell *182*, 226–244
July 9, 2020
https://doi.org/10.1016/j.cell.2020.06.012

CellPress

CellPress
OPEN ACCESS

Cell

Resource

Proteogenomics of Non-smoking Lung Cancer in East Asia Delineates Molecular Signatures of Pathogenesis and Progression

Yi-Ju Chen,[1,24] Theodoros I. Roumeliotis,[2,24] Ya-Hsuan Chang,[3,24] Ching-Tai Chen,[4,25] Chia-Li Han,[5,25,*] Miao-Hsia Lin,[1,25] Huei-Wen Chen,[6] Gee-Chen Chang,[7,8] Yih-Leong Chang,[9] Chen-Tu Wu,[9] Mong-Wei Lin,[10] Min-Shu Hsieh,[9] Yu-Tai Wang,[11] Yet-Ran Chen,[12] Inge Jonassen,[13] Fatemeh Zamanzad Ghavidel,[13] Ze-Shiang Lin,[14] Kuen-Tyng Lin,[1] Ching-Wen Chen,[14] Pei-Yuan Sheu,[14] Chen-Ting Hung,[14] Ke-Chieh Huang,[1] Hao-Chin Yang,[1] Pei-Yi Lin,[1] Ta-Chi Yen,[1] Yi-Wei Lin,[10] Jen-Hung Wang,[4] Lovely Raghav,[3,15] Chien-Yu Lin,[3] Yan-Si Chen,[3] Pei-Shan Wu,[1]

(Author list continued on next page)

[1]Institute of Chemistry, Academia Sinica, Taipei, Taiwan
[2]Functional Proteomics Group, Chester Beatty Laboratories, The Institute of Cancer Research, London SW3 6JB, UK
[3]Institute of Statistical Science, Academia Sinica, Taipei, Taiwan
[4]Institute of Information Science, Academia Sinica, Taipei, Taiwan
[5]Master Program in Clinical Pharmacogenomics and Pharmacoproteomics, College of Pharmacy, Taipei Medical University, Taipei, Taiwan
[6]Graduate Institute of Toxicology, College of Medicine, National Taiwan University, Taipei, Taiwan
[7]Division of Chest Medicine, Department of Internal Medicine, Taichung Veterans General Hospital, Taichung, Taiwan
[8]Faculty of Medicine, School of Medicine, National Yang-Ming University, Taipei, Taiwan
[9]Department of Pathology, National Taiwan University Hospital and National Taiwan University College of Medicine, Taipei, Taiwan
[10]Division of Thoracic Surgery, Department of Surgery, National Taiwan University Hospital, Taipei, Taiwan
[11]National Applied Research Laboratories, National Center for High-performance Computing, Hsinchu, Taiwan
[12]Agricultural Biotechnology Research Center, Academia Sinica, Taipei, Taiwan
[13]Computational Biology Unit (CBU), Informatics Department, University of Bergen, Bergen, Norway
[14]Department of Clinical Laboratory Sciences and Medical Biotechnology, College of Medicine, National Taiwan University, Taipei, Taiwan
[15]Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Bioinformatics Program, Taiwan International Graduate Program, Hsinchu, Taiwan

(Affiliations continued on next page)

SUMMARY

Lung cancer in East Asia is characterized by a high percentage of never-smokers, early onset and predominant *EGFR* mutations. To illuminate the molecular phenotype of this demographically distinct disease, we performed a deep comprehensive proteogenomic study on a prospectively collected cohort in Taiwan, representing early stage, predominantly female, non-smoking lung adenocarcinoma. Integrated genomic, proteomic, and phosphoproteomic analysis delineated the demographically distinct molecular attributes and hallmarks of tumor progression. Mutational signature analysis revealed age- and gender-related mutagenesis mechanisms, characterized by high prevalence of APOBEC mutational signature in younger females and over-representation of environmental carcinogen-like mutational signatures in older females. A proteomics-informed classification distinguished the clinical characteristics of early stage patients with *EGFR* mutations. Furthermore, integrated protein network analysis revealed the cellular remodeling underpinning clinical trajectories and nominated candidate biomarkers for patient stratification and therapeutic intervention. This multi-omic molecular architecture may help develop strategies for management of early stage never-smoker lung adenocarcinoma.

INTRODUCTION

Lung cancer remains the most common malignancy and the leading cause of cancer mortality worldwide (Bray et al., 2018) and has been mainly attributed to direct tobacco exposure (Bach, 2009). However, its incidence in never-smokers remains a significant health problem globally, especially in East Asia and most predominantly among women; the non-smoking-related etiology and carcinogenesis remain poorly understood (Jemal et al., 2018; Sun et al., 2007). In Taiwanese population, never-smoker patients are predominant (53%), especially among females (93%) (Tseng et al., 2019). Additionally, early onset is a distinct feature of lung adenocarcinoma (LUAD) in East Asia, particularly among never-smokers (Kawaguchi et al.,

CellPress
OPEN ACCESS

Chi-Ting Lai,[1] Shao-Hsing Weng,[1] Kang-Yi Su,[14,16] Wei-Hung Chang,[12] Pang-Yan Tsai,[12] Ana I. Robles,[17] Henry Rodriguez,[17] Yi-Jing Hsiao,[14] Wen-Hsin Chang,[18] Ting-Yi Sung,[4,*] Jin-Shing Chen,[19,*] Sung-Liang Yu,[14,16,*] Jyoti S. Choudhary,[2,*] Hsuan-Yu Chen,[3,20,*] Pan-Chyr Yang,[21,22,*] and Yu-Ju Chen[1,23,26,*]

[16]Department of Laboratory Medicine, National Taiwan University Hospital, Taipei, Taiwan
[17]Office of Cancer Clinical Proteomics Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA
[18]Institute of Molecular Medicine, National Taiwan University College of Medicine, Taipei, Taiwan
[19]Department of Surgery, National Taiwan University Hospital, Taipei, Taiwan
[20]Ph.D. Program in Microbial Genomics, National Chung Hsing University, Taichung, Taiwan
[21]Department of Internal Medicine, National Taiwan University, Taipei, Taiwan
[22]Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan
[23]Department of Chemistry, National Taiwan University, Taipei, Taiwan
[24]These authors contributed equally
[25]These authors contributed equally
[26]Lead Contact
*Correspondence: was@tmu.edu.tw (C.-L.H.), tsung@iis.sinica.edu.tw (T.-Y.S.), chenjs@ntu.edu.tw (J.-S.C.), slyu@ntu.edu.tw (S.-L.Y.), jyoti.choudhary@icr.ac.uk (J.S.C.), hychen@stat.sinica.edu.tw (H.-Y.C.), pcyang@ntu.edu.tw (P.-C.Y.), yujuchen@gate.sinica.edu.tw (Y.-J.C.)
https://doi.org/10.1016/j.cell.2020.06.012

2010). Genetic factors and exposure to environmental carcinogens may present risk factors contributing to these population differences (Samet et al., 2009). For instance, in Taiwan, air pollution has been shown to correlate with the incidence of lung cancer in never-smokers (Tseng et al., 2019). To complement the advances in precision therapy for advanced stage, early detection and prevention may create better clinical and economic benefits for patients and LUAD management. Thus, it is crucial to understand the early processes and progression of oncogenesis, as well as the contributing factors associated with endogenous and environmental mutagens underlying the unique characteristics of non-smoking LUAD in East Asia.

Significant unmet clinical needs remain in early stage LUAD. About 20% of stage I patients still relapse after surgical resection worldwide (Sawabata et al., 2010). At the molecular level, *EGFR* activating mutations, comprising mainly L858R mutation and the E746_A750 exon 19 deletion, occur much more frequently in East Asia (>50%, especially in never-smoker females) (Yang et al., 2020; Shi et al., 2014). Although patients bearing *EGFR* mutations benefit from targeted therapies using tyrosine kinase inhibitors, most of them eventually develop resistance (Tomasello et al., 2018). Distinctly, patients with *EGFR*-L858R mutation display shorter overall survival and a higher tendency to develop malignant pleural effusion and cancer metastasis compared to patients with *EGFR* exon 19 deletion (Kelly et al., 2018). A more comprehensive understanding of the molecular remodeling associated with oncogenic *EGFR* mutations in early stage will help to devise more effective therapeutic approaches.

The mutational spectrum of LUAD has been extensively explored by several genomic studies, mostly representing smoking-predominant cohorts (Campbell et al., 2016; Cancer Genome Atlas Research, 2014; Imielinski et al., 2012). These studies generated comprehensive catalogs of somatic mutations in Western populations and mutational subtypes associated with smoking. Multi-dimensional "omics" strategies encompassing proteome and phosphoproteome profiling of cancer tissues, in conjunction with genomic analysis, have elucidated new disease subtypes and signaling pathways, as well as potential targets for therapeutic development (Gao et al., 2019; Vasaikar et al., 2019; Zhang et al., 2016). Although the genomic profiles of lung cancer in Chinese patients were recently reported (Luo et al., 2018; Wang et al., 2018a; Zhang et al., 2019), a comprehensive proteogenomic profiling that can inform on the etiology and unique features of never-smoker and early onset of LUAD in East Asia is currently lacking.

In this study, we performed comprehensive genomic, transcriptomic, proteomic, and phosphorylation analysis of patient-matched early stage LUAD tumors, the predominant type of non-small cell lung cancer (NSCLC), and normal adjacent tissues (NATs) obtained from Taiwanese patients representative of the East Asian population. This integrated proteogenomic view revealed the molecular attributes associated with early events and non-smoking-related processes in LUAD, serving as a resource to the cancer community to further delineate the underlying biology and address the unmet clinical needs. Another large, deep scale proteogenomics study of lung adenocarcinoma in a geographically diverse set of patient samples appears in this issue (Gillette et al., 2020).

RESULTS

Proteogenomic Landscape of East Asian Lung Adenocarcinoma Highlights Demographic Differences and Progression Landmarks

To characterize the proteogenomic landscape of lung adenocarcinoma in East Asia, whole exome sequencing (WES), RNA-seq, proteomics, and phosphoproteomics data were collected from patient-matched tumor and NAT from 103 treatment-naive patients from Taiwan. The clinicopathological characteristics of patients and tumors are summarized in Table S1A. The prospective cohort consisted of 42% male and 58% female patients, 83% non-smokers, and had a median age of 63 years. This cohort (henceforth TW) is distinct from previous lung cancer genomics studies composed of more than 70% of smokers (Campbell et al., 2016; Cancer Genome Atlas Research, 2014; Imielinski et al., 2012). Histologically, 89% of the tumors were adenocarcinoma, and 80% were at early stages IA and IB (Table S1B). In the adenocarcinoma group (n = 91), a total of 23,145 nonsynonymous somatic single nucleotide variants (SNVs, Table S1C) were identified. At the transcriptional level, a total of 30,155 RNAs were quantified (Table S1D). Using isobaric labeling (Figure S1A), more than 10,000 unique proteins and 20,000

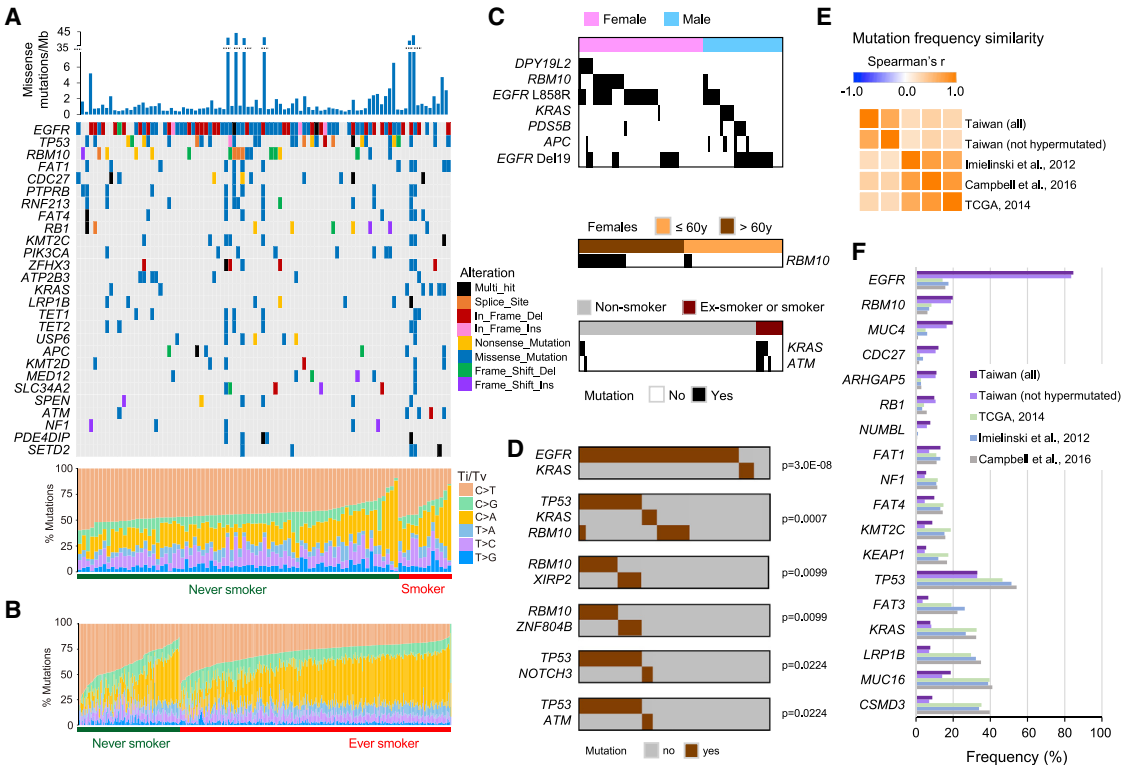


Figure 1. Genomic Landscape of the Asian LUAD Cohort

(A) Mutation colormap of common cancer genes and percentages of single nucleotide variants (SNVs) per patient in the Taiwan cohort.

(B) The percentages of SNVs in the TCGA cohort.

(C) Clustergrams of gender-enriched mutations (top panel), age-related *RBM10* mutations in females (middle panel), and smoking-associated mutations (bottom panel).

(D) Clustergrams of mutually exclusive mutated genes.

(E) Correlation plot of the mutation frequencies observed in the Taiwan cohort compared to previously published cohorts.

(F) Bar plots of mutational frequencies for genes with significant difference between the Taiwan cohort and previously published LUAD studies.

See also Figure S1.

phosphosites were quantified (Figures S1B–S1D; Table S1E–S1G). Two reference samples from a pool of tumor and normal tissues and a pool of late stage tumors that were included in all batches showed a mean correlation of 0.88 and 0.83 for the proteome and phosphoproteome respectively, confirming high technical reproducibility (Figures S1E and S1F).

Genomic profiles of genes implicated in cancer according to the Cancer Gene Census (COSMIC) are shown in Figure 1A and Table S1C. *EGFR* mutations occurred in most patients (85%) as expected, followed by mutations in *TP53* (33%) and *RBM10* (20%). The overall proportions of SNVs were different between TW and TCGA (the Cancer Genome Atlas) cohorts ($p = 0.0005$, Figure S1G), with cytosine to thymine (C>T) transition being the most frequent in the TW cohort (Figure 1A, bottom panel) and smoking-related cytosine to adenine (C>A) transversion being the most frequent in the TCGA cohort (Figure 1B). Non-smokers in the two cohorts showed similar proportions of C>T transitions (Table S1H). In contrast, the smoking related C>A transversions are significantly prominent in the TCGA cohort ($p = 0.0053$, Table S1I), especially in smokers ($p < 0.0001$). Most interestingly, no significant difference in C>A transversions was observed between smokers and non-smokers in the TW cohort (Table S1I). These observations suggest less significant smoking-related features, implicating other factors contributing to the genomic landscape of TW cohort.

Notably, *RBM10* and *EGFR*-L858R mutations were frequent in females, whereas *KRAS* and *APC* were often mutated in males (Fisher's exact test, $p < 0.05$; Figure 1C, top panel). *KRAS* and *ATM* were prominent mutations among patients with smoking history ($p < 0.01$; Figure 1C, bottom panel). Notably, *RBM10* mutations were more prevalent in older females (Figure 1C, middle panel) and coincided with downregulation of both RNA ($p = 0.021$) and protein ($p = 0.036$) levels, which was not significant in males (Figure S1H). The frequently mutated genes were tested for mutual exclusivity that may indicate novel synthetic lethality or distinct clonal evolution (Hua et al., 2016). In addition to the expected mutual exclusivity between *EGFR* and *KRAS* mutations ($p < 0.01$) (Suda et al., 2010), *RBM10* mutations were mutually exclusive with *TP53*, *KRAS*, *XIRP2*, and *ZNF804B* mutations ($p < 0.05$, Figure 1D). Correlation analysis across studies using mutation frequencies from cBioPortal

(Cerami et al., 2012) further reflects the distinct profile of our cohort (Figure 1E). The *EGFR* and *RBM10*, as well as two cell-cycle-related genes (*CDC27* and *RB1*) have much higher mutation frequency in our cohort, whereas somatic mutations in *TP53*, *KRAS*, and *KEAP1* were more prevalent in the other three studies (Figure 1F). Even comparing non-smokers in the TW and TCGA LUAD cohorts, several genes had significantly different mutation frequencies (Table S1J). For example, top-ranking genes *EGFR*, *RBM10*, and *RNF213* have significantly higher frequencies (3.7- to 5.9-fold) in TW cohort, while *KRAS* mutation occurs more frequently (4.5-fold) in TCGA cohort. Somatic mutations on *ATP2B3* and *TET2* also occur more frequently in TW cohort. The results indicate differences of cancer genomes for the never-smokers between TW and TCGA LUAD. RNA-seq, proteomics, and phosphoproteomics data were integrated to devise a multi-omics taxonomy. Principal-component analysis (PCA) using row-mean scaled data (log_2-transformed) showed a clear separation of the tumor and normal tissues at both the RNA and protein levels, as well as distinct clusters of the reference samples, confirming the absence of batch effects and revealing the higher variation of tumor compared to NAT (Figure 2A). The RNA-to-protein correlation using log_2T/N (tumor/normal) values was moderate to low with sample-wise and gene-wise median Spearman correlations of 0.31 and 0.14, respectively (Figures 2B and 2C; Table S2A). Only 22% proteins displayed significant positive correlations with the cognate RNA (Spearman, Benj. Hoch. false discovery rate [FDR] < 0.05; Figure 2C). Enrichment analysis showed a pathway-dependent RNA-to-protein correlation, with basic cellular functions poorly corresponding to RNA (Figure 2D; Table S2B). Additionally, pathway enrichment analysis using the protein median log_2T/N values across patients revealed the overall regulation trends (Figure 2E; Table S2C). Taken together, these analyses indicate transcriptionally modulated upregulation of DNA replication, glycolysis, glutathione metabolism, and immune-related pathways, while upregulation of DNA repair, protein processing and transport pathways, and downregulation of cell-adhesion-related pathways were more apparent at protein level. Focusing on the NSCLC pathway, most proteins and their phosphorylation sites were differentially regulated. Though protein expression of EGFR, ERBB2, Ras, and PKC were downregulated, many of their downstream signaling protein nodes such as JAK3, STATs, PI3K, AKT, MEK, EML4, PLCG2, and STK4 were consistently upregulated, likely mediated by phosphorylation of these kinases and known oncogenes (Figure 2F; Table S2D). The phosphorylation sites on the Raf/MEK/ERK axis displayed high inter-patient variation based on the standard deviation of the regressed phosphorylation values across patients, indicating patient-specific regulation of the MAPK pathway (Figure 2F; Table S2E).

To elucidate the molecular dynamics of tumor progression, we classified patients into three groups; IA, IB, and ≥II stages; and performed differential expression analysis at both protein and RNA levels using ANOVA. Differentially expressed proteins and RNAs ($p < 0.05$) were further divided into two clusters by k-means clustering, and enrichment analysis of biological terms and Gene Ontology biological process (GO-BP) was performed for the clusters with progressive up- or downregulation across stages (Figures 2G and 2H). Several key processes and terms such as DNA replication, canonical glycolysis, proteasome, antibacterial humoral response, glycosyltransferase, and actin filament organization were common between the two molecular levels (Figures 2I and 2J). Proteins that function in cell-to-cell communication, signaling, and plasma membrane such as integrins, G-protein coupled receptors, ion channels, adaptive immunity, and antigen presentation presented an overall negative regulation trend during progression (Figure 2I). In contrast, proteins in glycolysis, DNA replication, stress response, and protein processing, turnover, and trafficking processes were upregulated in the later stages. The upregulation of DNA replication and repair processes, as well as the loss of cilium assembly genes, were most prominent at the RNA level (Figure 2J). Lung cancer is a very heterogeneous disease at a cellular and histological level. Thus, it is noted that tumor heterogeneity may partially contribute to these differences. Nevertheless, these results highlight the importance of multi-omics integration to identify dysregulation of molecular homeostasis during tumor progression.

In summary, our results reveal a demographically distinct genomic landscape with different driver alteration frequencies. Its proteogenomic characterization shows that cellular transformation toward a more advanced cancer stage is characterized by an overall RNA-to-protein activation of replication with a parallel negative regulation of the proteome components involved in plasma membrane signaling and communication. Furthermore, the identified proteomic and RNA signatures represent the hallmarks of biological process remodeling that occurs during tumor progression.

The Impact of Genomic Alterations in the Proteome of NSCLC

Next, we delineated the direct and indirect consequences of genomic aberrations in our cohort at the transcriptome and proteome levels. Using customized protein databases incorporating somatic mutations of individual patients, 337 mutated peptides corresponding to 319 proteins were identified (q-value < 0.01, FDR < 1%). Among these, variant isoforms of 15 cancer driver genes were identified, such as *TP53BP1* D358E, *RNF213* E1272Q, and D1331G, and *KRAS* G12C mutations in the top-ranking genes (Figure S2; Table S2F). Truncating mutations in *RBM10* showed a systematic negative effect on both RNA and protein levels, whereas missense mutations in *KRAS, LABM1*, and *PIK3CA* were associated with increased protein expression only, possibly through increased stability (centered log_2T/N values, t test $p < 0.003$; Figure 3A; Tables S3A and S3B). Elevated phosphorylation of EGFR S1064 and Y1197 has been reported in response to EGF in lung cancer cells (Zhang et al., 2015). Although the impact of mutations in EGFR protein abundance was not conclusive, *EGFR* activating mutations (L858R and Del19) correlated with increased phosphorylation of S1064 and Y1197 (Figure 3B) (Tam et al., 2009), reflecting the activation of mutated EGFR in the patients. We also performed phospho-correlation analysis using the phosphorylation log_2T/N values normalized to the corresponding protein log_2T/N by linear regression and filtered by Spearman's p values. Downstream activation of the MAPK signaling can be evidenced by the positive correlation of EGFR-pY1197 with MAP2K2-pT394 (Koch et al., 2016), which further correlates with its substrate MAPK3-pT198/pT202 (Figure 3C), in turn with pMAPK1 and other downstream phosphoproteins

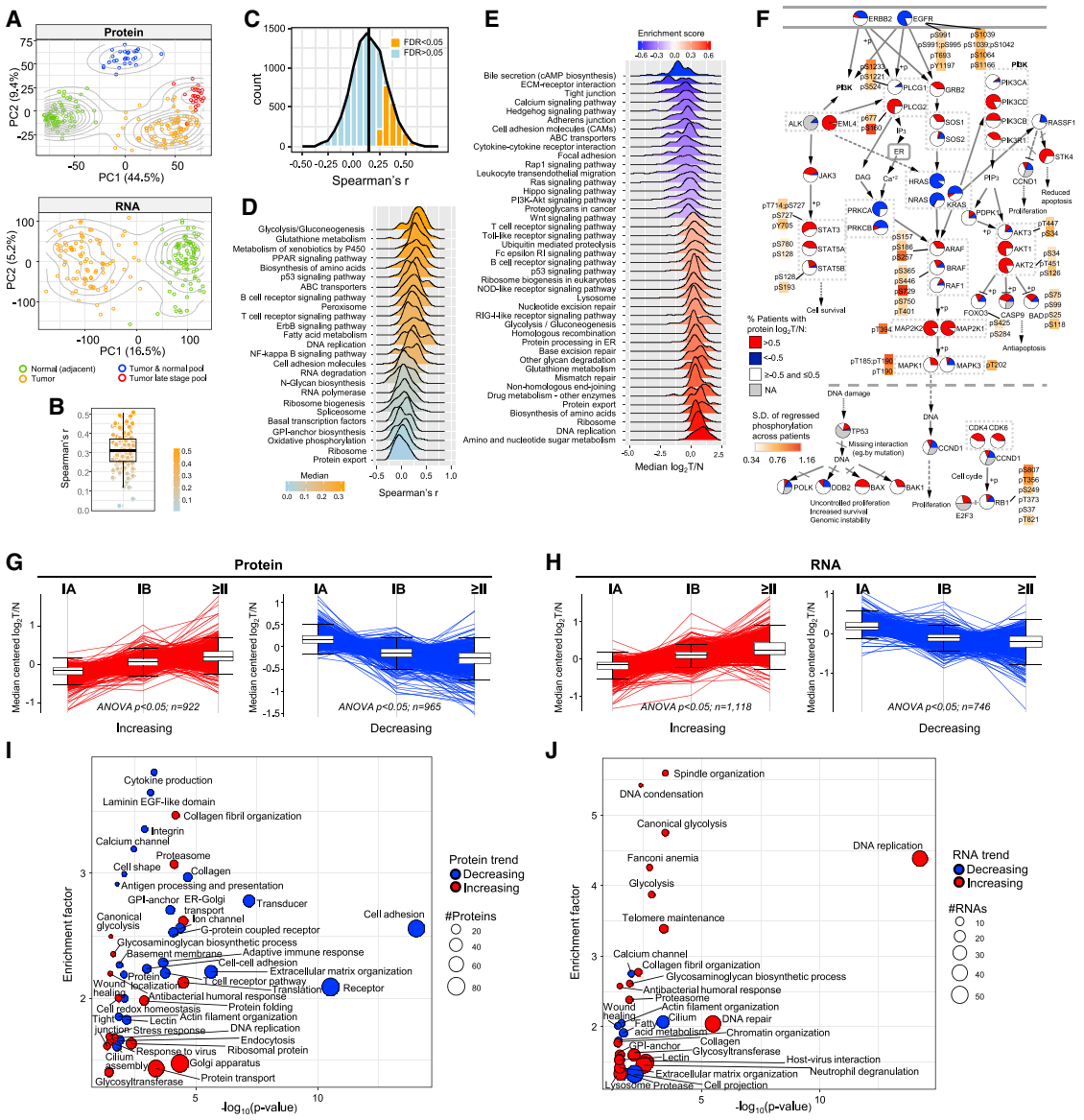


Figure 2. Molecular Variation, Proteogenomic Relationships, and Tumor Progression Hallmarks

(A) Principal-component analysis (PCA) of the protein and RNA $\log_2$-scaled values across the LUAD patients and reference samples.

(B) Boxplot of sample-wise RNA-to-protein correlations. Scale bar shows the Spearman correlation. In the box plots, the central line represents median, bounds of box represent the first and third quartiles, and the upper and lower whiskers extend to the highest or the smallest value within 1.5 × interquartile range (IQR).

(C and D) Histogram (C) and KEGG pathways enriched for higher or low gene-wise RNA-to-protein correlations (D, Benj. Hoch. FDR < 0.05).

(E) KEGG pathway enrichment analysis using the median $\log_2$T/N values of proteins (Benj. Hoch. FDR < 0.05).

(F) NSCLC pathway adapted from KEGG. The pie charts show the percentage of patients with up- or downregulation of the respective protein using $\log_2$T/N values. Selected phosphosites are shown and color coded according to their variation across patients using the standard deviation of regressed $\log_2$T/N phosphorylation values.

(G and H) Line plots and boxplots of selected gene sets with up- or downregulation trend during tumor progression at (G) protein level and (H) RNA level (ANOVA, $p < 0.05$).

(I and J) Scatterplot of significantly enriched UniProt keyword and Gene Ontology annotations with up- and downregulation trend based on their median abundance per pathology stage at (I) protein level and (J) RNA level.

See also Figure S2.

(RSK2, cPLA2, and STMN1, Figure S3A). Using their median relative abundance as a signature of MAPK pathway activity, patients were ranked from high to low MAPK signaling (Figure 3D). This indicated that the MAPK signaling pathway is commonly activated among both *EGFR*-WT (wild-type) and mutated patients with different degrees of activation. Patients without *EGFR* activating

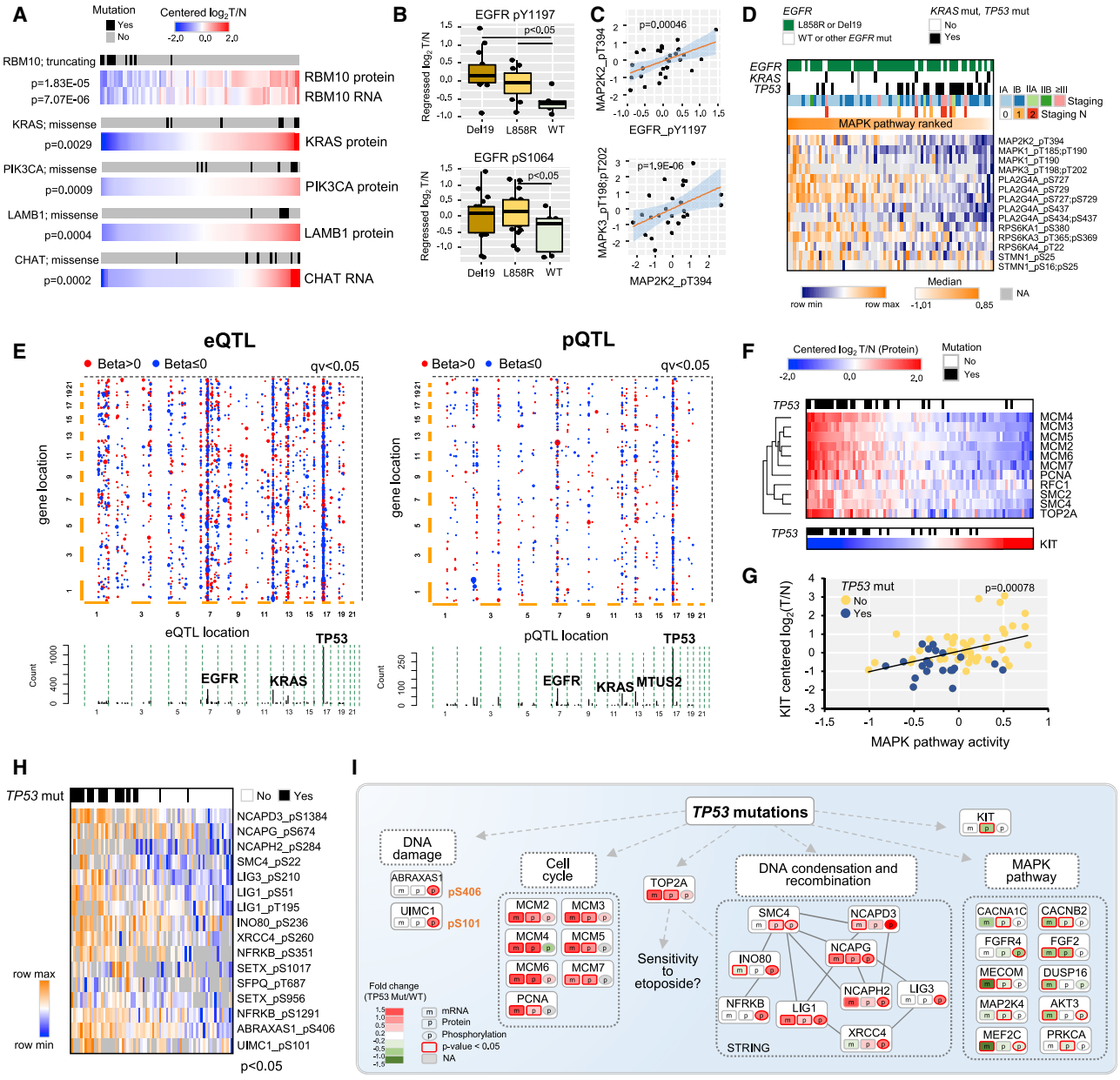


Figure 3. Impact of Mutations on the Proteome and Phosphoproteome of LUAD

(A) Heatmaps showing the direct effect of mutations on their encoded RNA and protein expression levels (centered log_2T/N).

(B) Boxplots illustrating the effect of *EGFR* activating mutations on EGFR phosphorylation (t-test, $p < 0.05$).

(C) Scatterplots of co-phosphorylation within the EGFR-MEK-ERK axis (Spearman's rank, $p < 0.05$).

(D) Ranked co-phosphorylation signature of the MAPK cascade aligned with clinical features.

(E) Two-dimensinal plot representing eQTLs and pQTLs with variants (x axis) and associated genes (y axis). The size of the points is increasing with the confidence of the association.

(F) Heatmap of the relative abundance of cell-cycle-related proteins (top panel) and KIT protein (bottom panel) that were significantly associated with *TP53* mutations.

(G) Scatterplot of the MAPK pathway score and KIT relative abundance across patients (Spearman's rank, $p = 0.00078$).

(H) Heatmap of phosphosites related to DNA condensation, recombination, and DNA damage response proteins positively associated with *TP53* mutations (t-test, $p < 0.05$).

(I) Summary of key *TP53* mutation associations.

See also Figure S3.

mutations frequently coincided with low MAPK signaling. Three of four *EGFR*-WT cases with higher MAPK activity harbored *KRAS* mutations (Figure 3D). It is noted that low MAPK signaling was observed for most tumors with *EGFR* mutations that also harbor *TP53* mutations (Figure 3D). Variation of MAPK pathway activity was observed within never-smokers with different *EGFR* activating mutations and is also influenced by *TP53* mutations (Figure S3B). Anti-correlation between *TP53* mutation and MAPK pathway activity has been observed in the TCGA cohort (Cancer Genome Atlas Research, 2014). Additionally, late stage tumors with lymph node metastasis showed lower MAPK activity (Figure 3D). Further studies are required to determine whether this dynamic MAPK pathway profile is associated with different clinical outcomes of these patients with *EGFR* activating mutation.

We also interrogated the indirect effects of somatic mutations at the RNA and protein levels by quantitative trait locus (QTL) analysis using centered $\log_2$T/N values (Table S3C and S3D). A total of 359 variant-RNA and 87 variant-protein interactions (FDR < 0.1) were identified, which indicated *TP53* locus on chromosome 17 as an eQTL and pQTL hotspot (Figures 3E and S3C). Both eQTLs and pQTLs showed a strong positive association of *TP53* mutations with cell cycle genes, including six subunits from the minichromosome maintenance complex (MCM) and TOP2A at the protein level (Figures 3F and S3D). Although the mechanism underpinning the TP53-MCM association is unclear, the modulation of MCM levels on chromatin by mutant TP53 has been reported (Qiu et al., 2017). TP53 deficiency can sensitize cells to Topoisomerase II inhibitors (Yeo et al., 2016); therefore, the observed *TP53*-TOP2A association may reflect opportunity for synthetic lethality in NSCLC. In line with this hypothesis, using public drug response data, we found that lung cancer cell lines bearing *TP53* mutations were more sensitive to etoposide (Wilcoxon test, p = 0.021; Figure S3E) (Corsello et al., 2019; Ghandi et al., 2019). A pQTL between *TP53* mutations and lower abundance of the KIT oncogene could potentially explain the low MAPK activity (Figure 3F, bottom panel, and Figure 3G) (Du and Lovly, 2018). Additionally, *TP53* mutations were positively associated with higher phosphorylation of proteins involved in DNA condensation and recombination and DNA damage proteins (p < 0.05, Figure 3H), representing potential therapeutic targets (Tomkinson et al., 2013; Wang et al., 2018b). Higher phosphorylation of ABRAXAS1-pS406 and UIMC1-pS101 (Wang et al., 2007; Kim et al., 2007) may indicate a higher degree of DNA damage in the *TP53* mutant tumors. A summary of the *TP53* associations that may guide future studies and therapeutic opportunities is shown in Figure 3I and Tables S3E and S3F. In summary, our results show coordinated protein phosphorylation within the MAPK cascade, which is partially explained by *EGFR* and *KRAS* mutations and negatively correlates with tumor staging and *TP53* mutations. *TP53* mutations were also linked to DNA replication, control of DNA topological states, and response to DNA damage.

Mutational Profiles Associated with Endogenous and Environmental Mutagens and Proteogenomic Impact

The frequencies of mutated trinucleotide sequence motifs were analyzed using non-negative matrix factorization (NMF) (Brunet et al., 2004; Lee and Seung, 1999), and five mutational profiles were identified (Figure 4A). To reveal the potential contribution of endogenous and exogenous mutagens in these profiles, cosine similarity analysis against mutational signatures in human cancer (Alexandrov et al., 2013a, 2013b; Mayakonda et al., 2018) and environmental agents *in vitro* (Kucab et al., 2019) was performed. For comparison, the same enrichment analysis was performed in the TCGA cohort, where three mutational profiles were identified (Figure S4).

The mutational signatures best matching to those in the TW cohort were (1) deamination of 5-methylcytosine, (2) APOBEC cytidine deaminase, (3) exposure to tobacco mutagens from COSMIC, and (4) dibenz[a,j]acridine (DBAC) from Kucab et al. These mutational profiles also mapped to the signatures of (1) alkylating agent dimethylhydrazine (DMH), (2) simulated solar radiation (SSR), and (3) 1,8-dinitropyrene (1,8-DNP) with lower scores (Figure 4A; Table S4A). The APOBEC and tobacco COSMIC signatures, as well as the signatures of dibenzanthracenes (DBA), DMH, and SSR, were also obtained in the three mutational profiles of the TCGA cohort (Figure S4). DBAC and DBA are polycyclic aromatic hydrocarbons (PAHs) produced by the incomplete burning of organic matter. More importantly, 1,8-DNP (Nitro-PAH) found in particulate emissions from combustion products (https://pubchem.ncbi.nlm.nih.gov), and N-Nitrosopyrrolidine (a nitrosamine, with lower similarity score), commonly derived from tobacco, food, and drink (Gushgari and Halden, 2018), were uniquely enriched in the TW cohort (Figure 4A). The results are in line with the previous epidemiological evidence on the high regional exposure of these two categories of carcinogens in Asia (Jakszyn and Gonzalez, 2006; Yan et al., 2019).

The APOBEC mutation signature is attributed to the polynucleotide cytosine deaminases protein family (apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like). The APOBEC signature was based on C>T and C>G mutations surrounding *TCW* sequence associated with various cancer types (Alexandrov et al., 2013a; Cho et al., 2018; Roberts et al., 2013). To relate APOBEC signature with clinical features, patients were dichotomized into APOBEC-high and -low signature groups based on enrichment scores (Roberts et al., 2013). The analysis revealed that 44% of the patients had high APOBEC signature and over-representation in females (57%) compared to males (25%) (p = 0.0045, Table S4B). Based on the clinico-demographics of >50,000 lung cancer patients in the Taiwan Cancer Registry from 2011–2015, a trend of early onset is seen in the bimodal distribution of younger (peak at 58–61 years) and older patients (peak at 70–76 years). Thus, 60 year was used as a threshold for age classification in this study. It is noted that females with high APOBEC signature and WT *EGFR* were younger compared to females with low APOBEC signature and *EGFR* activating mutations (mean age 52.9 and 66.3 years, respectively, p = 0.038, Figures 4B and 4C). Specifically, APOBEC-high signature was predominantly present in 74% younger females ($\leq$60 years) and in all females without *EGFR* mutation, whereas no similar trend was observed in male patients (Figure 4D; Table S4C). Notably, neither gender- nor age-specific differences were observed in the TCGA cohort (Table S4D). These results suggest the potential and unique contribution of APOBEC mutagenesis to the early onset of non-smoking LUAD in females. Survival analysis on the TCGA cohort revealed that low APOBEC signature is associated with prolonged overall

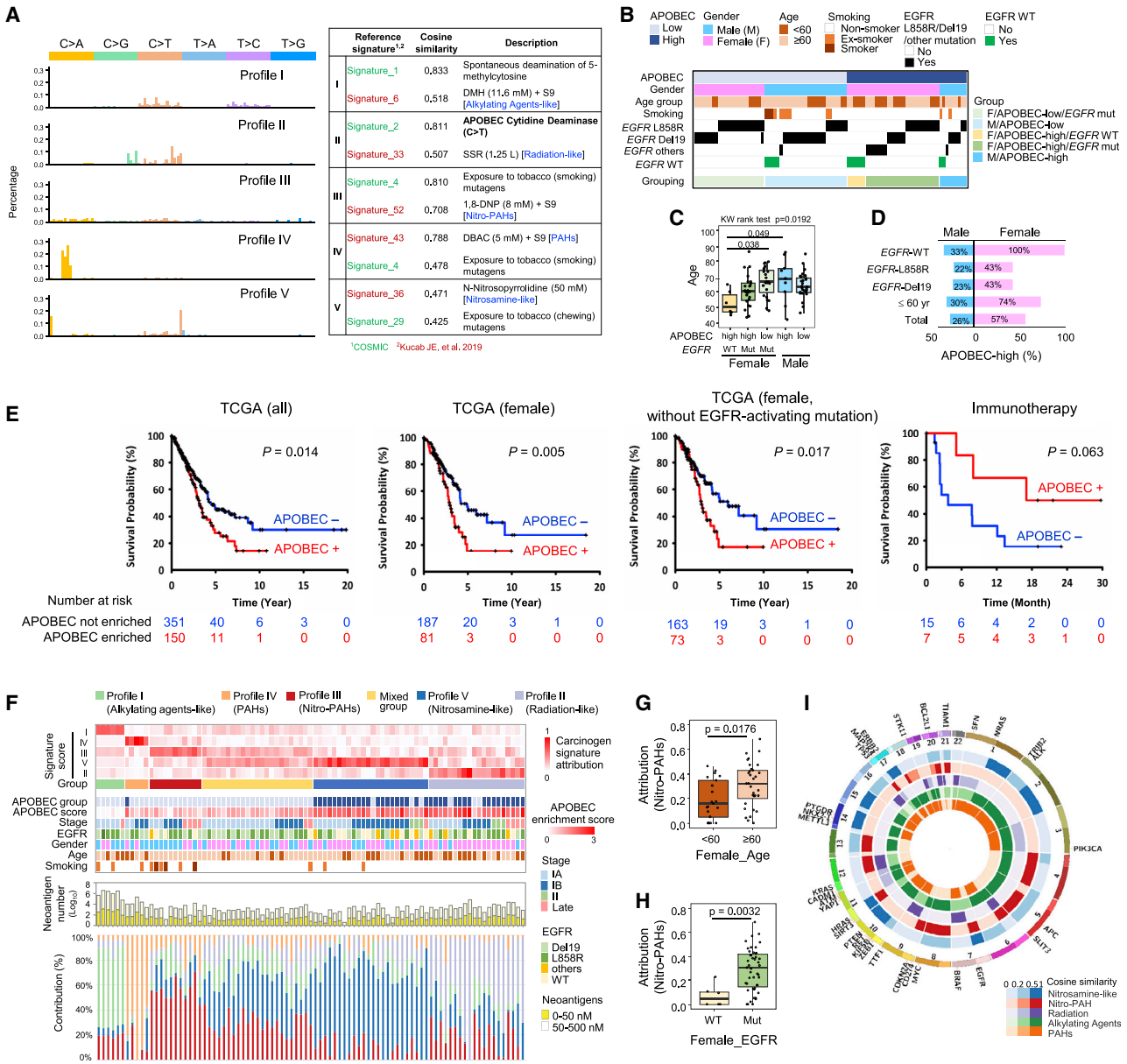


Figure 4. Identification of APOBEC and Carcinogen Mutational Signatures in Taiwan LUAD Cohort

(A) Trinucleotide motif frequency plots and enriched mutational signatures for five mutational profiles identified in TW cohort.

(B) Categorization of patients into five groups based on APOBEC signature, gender, and *EGFR* mutation status.

(C) Boxplots showing the age differences between the five groups (Kruskal-Wallis test, p = 0.0192, and pairwise comparisons of post-hoc analysis, p = 0.038 and p = 0.049).

(D) Percentages of patients with high APOBEC signature in male and female groups.

(E) Survival analysis for patients in APOBEC-high and -low signature groups in TCGA and immunotherapy cohorts (log-rank test).

(F) The relative percentage of each mutational signature profile and the corresponding clinical features of individual patients. Bar-chart presenting the number of predicted neoantigen for each patient.

(G and H) Boxplots showing the attribution of nitro-PAHs signature for (G) age and (H) *EGFR* mutation status of female patients (Wilcoxon rank-sum test, p = 0.0176 and p = 0.0032, respectively).

(I) Chromosome view visualizing the localization of key oncogenes or tumor suppressors (outer ring) and neighboring regions enriched among the five carcinogen signatures (inner rings).

See also Figure S4.

survival in all patients ($p = 0.014$), all females ($p = 0.005$), and females without *EGFR*-activating mutations ($p = 0.017$, Figure 4E). Notably, high APOBEC signature was associated with a marginally significant prolonged progression-free survival for an advanced NSCLC cohort treated with combination immunotherapy (PD-1 and CTLA-4; Figure 4E) (Hellmann et al., 2018). Despite the promise of immunotherapy, patients with *EGFR* or *ALK* mutation have poor response compared to WT patients (Mhanna et al., 2019). The *EGFR*-WT female group in our cohort has APOBEC-high signature, which may be in line with current knowledge that *EGFR*-WT patients are better candidates for immunotherapy. Our findings suggest that high APOBEC signature may help identify patients, such as young *EGFR*-WT female patients that are anticipated to respond to immunotherapy.

We further aligned patient characteristics with mutational signature groups, including prediction of tumor antigen load. Tumor antigens play an important role in T-cell-mediated antitumor immunity (Jiang et al., 2019). To explore the neoantigen landscape in our cohort originating from somatic mutations, we performed neoantigen load analysis using POLYSOLVER and NetMHC (v4.0; Andreatta and Nielsen, 2016). The numbers of neoantigens were variable across patients (Figure 4F). Comparing the APOBEC and carcinogen-like signatures, tumors enriched for alkylating agents-like signature showed more neoantigens (Kruskal-Wallis rank sum test, $p < 0.0001$). These may be referred to as hot tumors, since they have more neoantigens likely to be recognized by T cells and are more amenable to immunotherapies. The percentages of mutational profiles IV, III, and I matching the PAHs, Nitro-PAHs, and Alkylating agents, respectively, were higher in tumors from patients with smoking history (Figure 4F). Conversely, tumors predominantly with mutational profiles II (radiation-like) and V (nitrosamine-like) positively correlated with APOBEC signature. A "mixed group" (25.8% of all patients) with undefined signature was composed of never-smokers without APOBEC signature, whereas 95% of the patients in this group had *EGFR* activating mutation (Table S4E). Notably, in female patients, the distribution of mutation profile III (Nitro-PAHs signature) was overrepresented in the older group ($p = 0.0176$, Figure 4G), as well as in females with *EGFR* activating mutations ($p = 0.0032$, Figure 4H).

To identify chromosomal regions that may be more susceptible to mutagenesis by the five environmental factors, chromosome-wise enrichment analysis (Mayakonda et al., 2018) was performed by overlaying the somatic mutations on the carcinogen signatures. Results showed that chromosomal regions enriched for the carcinogen signatures overlapped with the genomic locations of oncogenes or tumor suppressor genes such as *TP53*, *ERBB2*, *MYC*, and *APC* (Figure 4I). For example, both Nitro-PAHs and Nitrosamine-like signatures were enriched in the chromosome 7p, which is the broader genomic location of *EGFR* gene. The regions of *TP53* (Chr.17p) and *RB1* (Chr.13q) were enriched in nitro-PAHs signature, whereas chromosome 12p, where *KRAS* is located, was enriched with PAHs and alkylating agent signatures (Figure 4I). Whether these patterns represent susceptibility to higher frequency of driver mutations in East Asia remains to be investigated.

Next, we interrogated the source and consequences of the APOBEC mutational signature at the proteome and phosphoproteome levels. At protein level, six members of the APOBEC3 family, reported to associate with APOBEC mutagenesis (Roper et al., 2019), were identified in at least 30% of the patients. These were frequently upregulated in the tumors (Figure 5A) and showed greater upregulation in female compared to male patients, although this difference was not recapitulated at the RNA level (Figure 5B). APOBEC3F and APOBEC3G showed the most significant gender-specific differences ($p < 0.05$). Pathway analysis in female patients revealed a number of proteins involved in DNA repair and replication more abundant in the tumors with high APOBEC signature (Figures 5C and S5A; Table S5A). In addition, relative $\log_2$T/N values of these proteins showed positive correlation with the APOBEC3 members, especially with 3B, 3D, 3F, and 3G (Figure 5C). Higher expression of base excision repair (BER) proteins in APOBEC-high females, including MBD4, APEX1, FEN1, and POLD2, implicates a role of BER in counteracting APOBEC-induced mutagenesis. Phosphosites on proteins in DNA damage and repair processes, such as ATR-T1989 (Liu et al., 2011) and UIMC1-S101 (Kim et al., 2007), were also differentially regulated between the APOBEC-high and -low groups (ANOVA, $p < 0.05$, Figure S5B). Kinase enrichment analysis (Weidner et al., 2014) identified AurB, CK2, CDK1, and CDK2 as the top-ranking activated kinases in the APOBEC-high female group (Figure 5D; Table S5B). The activation of CDK1, CDK2, and AurB offers actionable intervention candidates for female patients with high APOBEC signature (Lin et al., 2018; Maslyk et al., 2017; Mross et al., 2016).

To assess the functional impact of carcinogen signatures, we performed 1D-annotation pathway enrichment analysis (Cox and Mann, 2012) using the $\log_2$T/N values of mRNA, proteomic, and phosphoproteomic data of individual patients (Benj. Hoch. FDR < 0.05). The mean values of omics data within each enriched pathway from individual patients was compared among different carcinogen groups (Figure 5E; Table S5C). Tumors harboring PAH or nitro-PAH signatures showed significant enrichment for pathways associated with metabolism and detoxification of chemical carcinogens (Kruskal-Wallis rank test, $p < 0.05$, Figure 5E), including the AHR and Cytochrome P450 pathways, known to contribute to carcinogenesis by PAH (Moorthy et al., 2015). The nitro-PAH and nitrosamines-like groups were dominated by DNA repair, ERBB/MAPK pathway, and TLR/RIG-1 T cell signaling, which potentially link to the tumor initiation, cell proliferation, EMT malignant progression, and immune modulation in early carcinogenesis (Figures 5E and 5F; Table S5D) (Moorthy et al., 2015).

Our findings implicate a potential role of the APOBEC signature in the manifestation of lung cancer at an early age in never-smoker females, possibly influencing disease outcome. Additionally, signatures akin to those caused by exogenous carcinogens *in vitro* and relevant carcinogenesis and oncogenesis pathways were observed predominantly in older females. Taken together, these findings uncover distinct age-related mutagenesis mechanisms in our female patients.

Proteomic Subtypes Resolve the Heterogeneity in Early Stage Lung Adenocarcinoma

Unsupervised consensus clustering (Wilkerson and Hayes, 2010) was used to classify the adenocarcinoma patients of

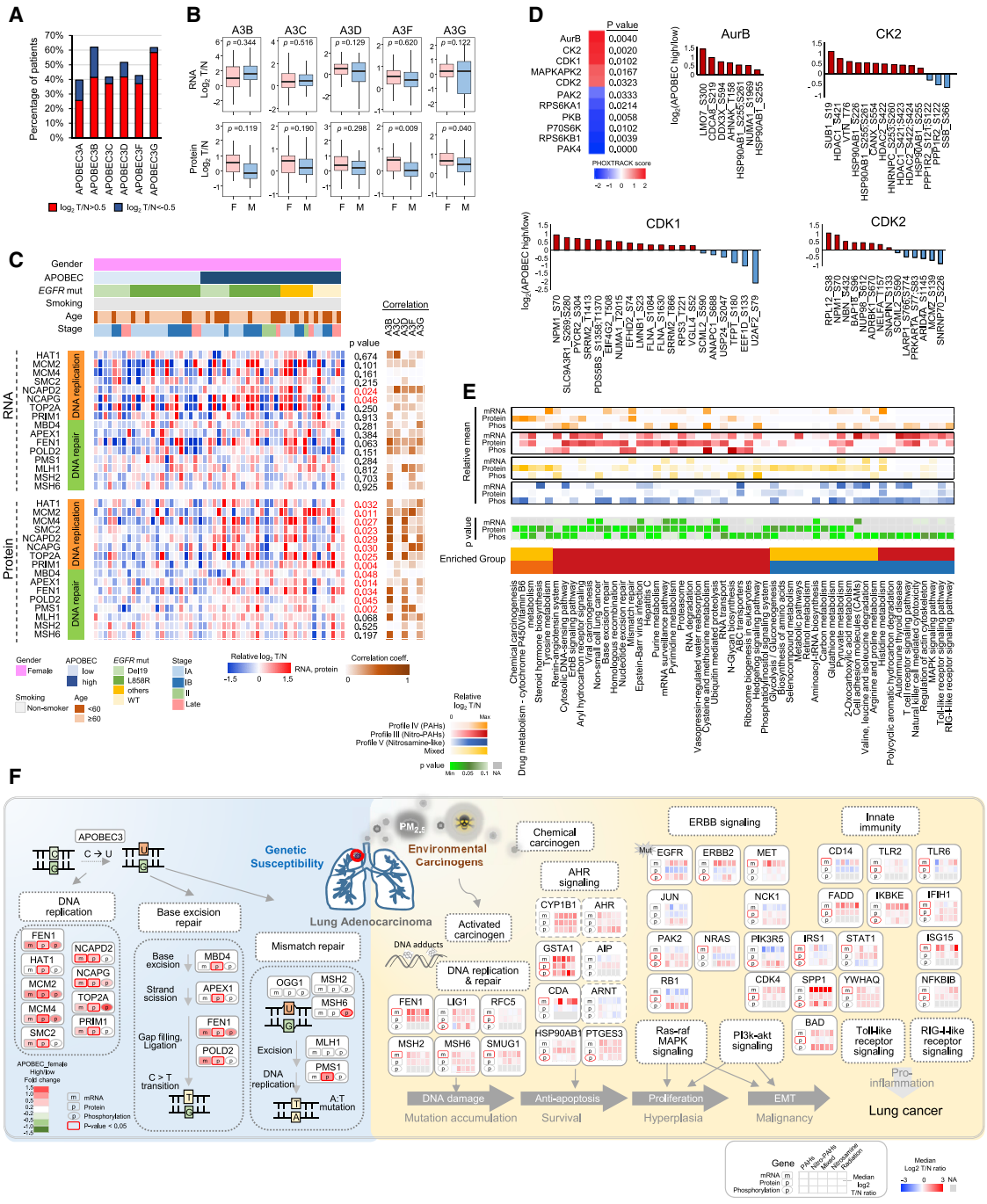


Figure 5. Altered Signaling Pathway Associated with APOBEC and Carcinogen Signatures

(A) Percentage of patients with up- or downregulated (red and blue respectively) protein abundance of APOBEC enzymes.

(B) Expression comparison of APOBEC family at mRNA and protein levels in female and male patients (Welch's t-test and pairwise comparison, $p < 0.05$).

(C) Heatmap showing relative abundance and Spearman's correlation between APOBEC family and DNA repair and replication molecules at mRNA and protein levels in females (Welch's t-test and pairwise comparison, $p < 0.05$).

(D) Kinase enrichment using the regulated phosphosites between APOBEC-high and -low groups and substrates of kinases with predicted activation (ANOVA, $p < 0.05$).

(E) Pathway enrichment analysis using the $\log_2$T/N values of mRNA, proteomic and phosphoproteomics data (Benj. Hoch. FDR < 0.05) of individual patients. The mean $\log_2$T/N values of each pathway were color coded (Kruskal-Wallis rank test, $p < 0.05$).

(F) Overview of significantly enriched pathways associated with APOBEC and carcinogen signatures. The $\log_2$T/N ratio was indicated either between APOBEC-high and -low female cohort or among five mutational signatures.

See also Figure S5.

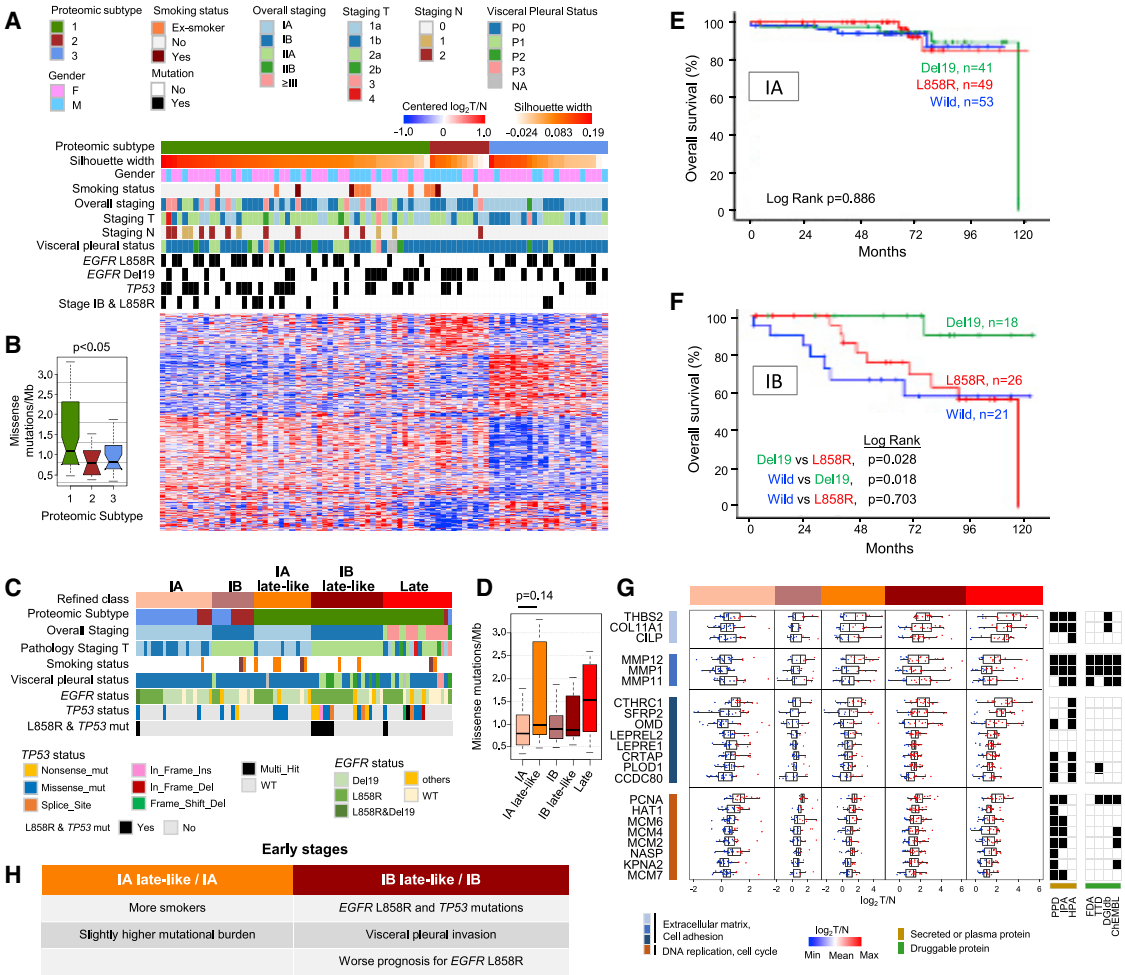


Figure 6. Proteomic Subtypes of the East Asian LUAD Cohort

(A) Heatmap of differentially regulated proteins among the three proteomic subtypes (ANOVA, FDR < 0.05) annotated with clinical features.

(B) Boxplot of the tumor mutational burden per proteomic subtype (Wilcoxon rank-sum test, $p < 0.05$).

(C) Refined staging classification informed by proteomics. Clinical annotations are color coded as in (A).

(D) Boxplot of the mutational burden per refined staging classification (Wilcoxon rank-sum test, $p < 0.05$).

(E and F) Survival plots in IA (E) and IB (F) patient groups with wild-type *EGFR* and with *EGFR*-L858R or Del19 mutations in an independent retrospective cohort of 209 patients (log-rank test).

(G) Boxplots for a panel of proteins with differential regulation between the refined classes (ANOVA, FDR < 0.05). The central line represents median, bounds of box represent the first and third quartiles, and the upper and lower whiskers extend to the highest or the smallest value within 1.5 × IQR. Their presence in serum is annotated from Plasma Proteome Database (PPD), Human Proteome Atlas (HPA), Ingenuity Pathway Analysis (IPA), and drug targets are annotated from FDA-approved drug targets, Therapeutic Target Database (TTD), The Drug-Gene Interaction database (DGIdb), and ChEMBL.

(H) Summary of the characteristics of the "late-like" class in stages IA and IB.

See also Figure S6.

our cohort into molecular subtypes. We identified three proteomic, three RNA, and four phosphoproteomic subtypes after excluding small or unstable clusters (Figure S6A). In addition to the inherent differences in the multi-omics profiles contributing to the subtyping results, it is noted that variations in cancer cellularity in tumors for the multi-omics analysis cannot be excluded. Alignment of the proteomic subtypes with clinical features revealed a strong separation by tumor staging, as well as by driver mutations (Figure 6A; Tables S6A and S6B). Subtype 1 comprised the largest group, over-representing late stage tumors (≥II), tumors with visceral pleural invasion, *TP53* mutations, and the highest mutational burden (Figure 6B). Subtype 2 is a smaller group consisting of early stage patients (mostly IA and IB) without *EGFR*-L858R mutations. Subtype 3 showed an over-representation of early stage (stage IA) largely lacking *TP53* mutations. We identified 1,514 differentially regulated phosphosites between the subtypes (regressed $\log_2$T/N values, ANOVA, FDR < 0.1) showing enrichment for pathways in cancer, PI3K-AKT signaling pathway, and cell cycle (Kyoto Encyclopedia of Genes and Genomes [KEGG], Fisher's exact test, Benj. Hoch. FDR < 0.1, Figure S6B). Interestingly, phosphosites from key pathways showed an overall higher

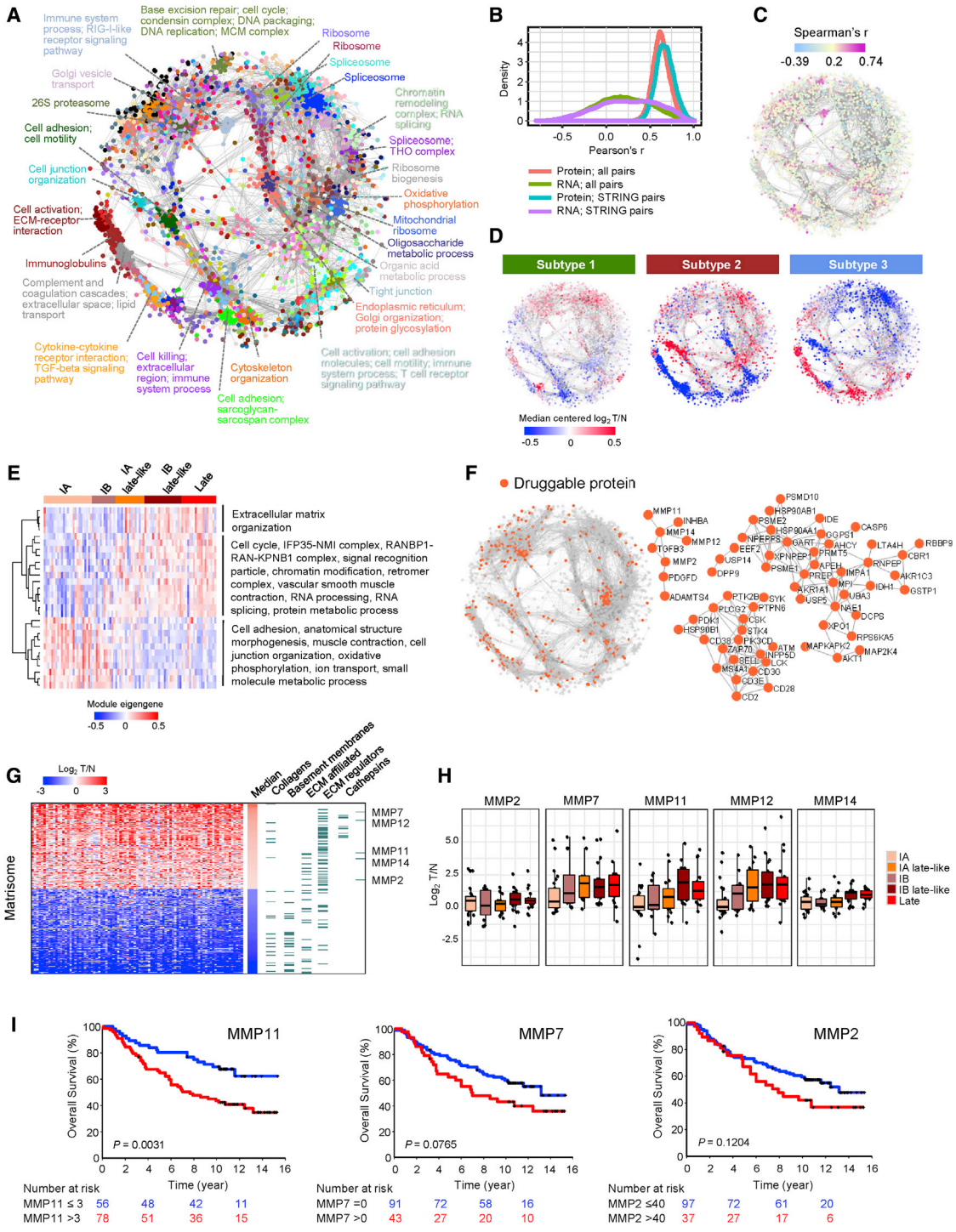


Figure 7. Weighted Gene Co-expression Network Analysis (WGCNA) and Association of Matrix Metalloproteinase Expression with Clinical Outcome

(A) Protein correlation network of 3,014 nodes and 44,665 edges. Nodes are color coded according to module membership. Representative enriched biological terms are shown for distinct modules.

(B) Density plots of the pairwise protein-protein and RNA-RNA correlations for the interactions shown in the network of (A).

(C) Overlay of the RNA-to-protein correlation on the network nodes.

(D) Overlay of the median-centered log_2T/N values per proteomic subtype on the network.

(E) Heatmap of differentially regulated modules between the refined classes.

(legend continued on next page)

phosphorylation in subtype 3 (Figure S6B), which would implicate a signaling signature that is activated early during tumor development. Specific examples are shown in Figure S6C, and druggable protein hits are denoted. Additionally, tyrosine phosphorylation on proteins involved in immune signaling were also in higher levels in subtype 3, suggesting enhanced regulation of immune signaling in the early stages (Figure S6D). Grouping by stage was less evident in the RNA clustering, although features such as poor pathological differentiation and angiolymphatic invasion were correlated more significantly (Figure S6E). At the phosphoproteome, we found a good separation by the APOBEC signature (phosphorylation subtypes 1 and 2, Figure S6E), possibly reflecting widespread alterations in signaling. Overall, the RNA and phosphorylation subtypes showed only partial overlap with the proteomic subtypes (Figure S6F), stressing the complementarity of multi-omic tumor profiling in patient classification efforts.

Notably, the proteome-based classification revealed that a number of early stage tumors were clustered together with late stage tumors in subtype 1 (Figure 6A). We propose that proteomic profiles could be leveraged to inform pathological staging and devised a refined classification representing the stage IA and IB patients found in proteomic subtype 1 as distinct classes; stage IA "late-like" and stage IB "late-like," respectively (Figure 6C). Compared to stage IA tumors, stage IA late-like tumors were enriched for patients with smoking history and slightly higher mutational burden (Figure 6D). Almost all stage IB patients with visceral pleural invasion (P1-3) were in the stage IB late-like class. Most importantly, at stage IB, tumors with *EGFR*-L858R mutation were prominently clustered into the proteomically annotated late-like group compared to stage IB tumors without this driver mutation. Additionally, the stage IB late-like class was enriched for *TP53* mutations compared to the stage IB tumors found in proteomic subtypes 2 and 3 (Figure 6C). Specifically, all stage IB patients with dual *EGFR*-L858R/*TP53* mutant tumors are found in the late-like class (Figure 6C).

To evaluate whether the refined classification can have clinical implications, we examined whether patients with *EGFR*-L858R have different outcome compared to the *EGFR*-Del19 patients in an independent retrospective cohort encompassing treatment-naive, completely resected pathologic stage IA (n = 143) and IB (n = 65) LUAD patients, comprising predominantly never-smoker (74%) and female (61%) patients, with survival data (Table S6C and S6D). Patients diagnosed at stage IA show no difference in overall survival between the L858R and Del19 *EGFR* mutation groups (Figure 6E). However, at stage IB, patients with *EGFR*-L858R had a significantly inferior overall survival compared to the Del19 patients (p = 0.028, Figure 6F). These results confirm that the proteomics-based five-stage molecular classification distinguishes the diverse clinical trajectories of patients with *EGFR*-L858R, likely diverging at stage IB. This is consistent with their reported higher tendency of cancer metastasis and shorter overall survival compared to Del19 mutations (Tsai et al., 2018; Wu et al., 2013).

Lastly, using differential expression analysis, we defined a panel of biomarker candidates to discriminate early stage from early late-like or late stage tumors (log_2T/N values, ANOVA, FDR < 0.05, Figure 6G). To explore their potential as serum biomarkers or therapeutic targets, we provide annotations from various resources including Plasma Proteome Database (PPD), Human Proteome Atlas (HPA), Ingenuity Pathway Analysis (IPA), FDA-approved drug targets, Therapeutic Target Database (TTD), Drug-Gene Interaction database (DGIdb), and ChEMBL. A summary of the characteristics of the late-like class is shown in Figure 6H. We propose that proteomics-based classification can categorize early stage NSCLC patients into groups associated with clinical outcome, beyond the level of clinical staging and genomic driver mutations.

Protein Network Characterization of Proteomic Subtypes Identify Candidate Biomarkers and Druggable Targets

To explore the biological characteristics of our cohort in an unbiased proteome-wide manner (Lapek et al., 2017; Roumeliotis et al., 2017; Ryan et al., 2017), weighted correlation network analysis (WGCNA) (Langfelder and Horvath, 2008) was performed using the centered log_2T/N values of 9,072 proteins (Table S7A). We distinguished 279 modules of which 195 showed significant enrichment for Gene Ontology, KEGG, and CORUM (the comprehensive resource of mammalian protein complexes) annotations (Fisher's exact test, Benj. Hoch. FDR < 0.05, Table S7B). We applied filters to trim the input and constructed a network of 3,014 nodes and 44,665 edges (Figure 7A; Table S7A). The mean Pearson correlation of all edges was 0.63, and 28% were known STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) interactions (Szklarczyk et al., 2019) with score >0.4. Notably, the respective correlations at the RNA level were low with a broad distribution (median 0.19), even when the comparison was restricted to known associations (median 0.25, Figure 7B), indicating that this network is a distinct attribute of the proteome. The RNA-to-protein correlation is shown in Figure 7C and depicts nodes with tighter transcriptional regulation. Superimposing the relative protein abundances per proteomic subtype on the network highlights functional differences (Figure 7D).

Using the module eigengenes, 29 smaller subnetworks were identified with differential regulation between the three proteomic subtypes (Figure S7A). A gradual increase in abundance of proteasome, cell cycle, and endoplasmic reticulum networks

(F) Protein nodes with known drug inhibitor (source: DGIdb) are highlighted on the network with orange. Druggable protein subnetworks with median log_2T/N >0.5 are magnified.
(G) Heatmap of the top differentially regulated matrisome proteins with protein-type annotations.
(H) Boxplots illustrating the abundances of MMP2, MMP7, MMP11, MMP12, and MMP14 in the different refined classes. The central line represents median, bounds of box represent the first and third quartiles, and the upper and lower whiskers extend to the highest or the smallest value within 1.5 × IQR.
(I) Kaplan-Meier plots for MMP11, MMP7, and MMP2 from immunohistochemistry staining data.
See also Figure S7.

was observed from subtypes 3 and 2 to subtype 1, comprising advanced stage tumors. Subtype 2 had a distinct profile with higher abundance of peroxisome, cell junction, adhesion, signal transduction, aminoacyl-tRNA biosynthesis, TRAPP complex, and chromatin modification networks. Signatures with higher levels in subtype 3 included cytokine-mediated signaling network, cilium, cell adhesion and extracellular matrix organization, cell junction, and endocytosis (networks 17–21 in Figure S7A). A significant number of protein modules were also differentially regulated between the refined classes (Figure 7E). We interrogated the networks to identify subnetworks that distinguish stage IB tumors from the stage IB late-like tumors (Figure S7A), highlighting molecules with roles in communication and microenvironment modulation, as well as a variety of other functions.

We found that two modules include immune-cell-specific proteins such as the transcription factors IKZF1, IKZF3, NFATC1, and NFATC2 (networks 25 and 26 in Figure S7A) and have higher expression in subtype 3, representing early stage tumors. Indeed, both modules contained proteins with predominantly strong expression in B cells, T cells, and natural killer (NK) cells, according to the Human Proteome Map database (Figure S7B) (Kim et al., 2014). To investigate this observation, CIBERSORT analysis on RNA-seq data for all tissues (Newman et al., 2015) was performed. Tumors separated from NATs with several immune cell types co-existing in about half of the tumors and several tumors presenting low levels of immune infiltration (Figure S7C). The protein network and RNA-based approaches were significantly correlated (Figure S7D), distinguishing patients with high or low immune-cell-related protein abundances (Figure S7E). Interestingly, patients with primary tumor located at the right side of lung present higher immune expression compared to those at the left side ($p = 0.0379$). Antigen processing and presentation (MHC class II) was also higher in subtype 3 (network 27). It remains to be investigated whether the immune infiltration correlates with patient outcome or response to immunotherapy. Given the important role of tumor microenvironment (TME) in tumor development and progression, we further used the ESTIMATE algorithm (Yoshihara et al., 2013) to deconvolute the contribution of stromal cells in the tumors using the RNA-seq data (Table S7C). To identify potential protein biomarkers for immunohistochemical validation of the TME, patients were classified into two groups with high and low stromal scores (<first and >third quartile), and differential protein expression was computed (Figure S7G). Among the top hits, CILP was identified with 2.6-fold higher abundance in the high stroma group (-ANOVA, $p = 0.00396$, Figure S7H). The CILP protein was identified with an average of 31 peptides and was mapped to a module associated with extracellular matrix organization, representing a promising stromal cell target candidate for further development using immunohistochemistry or mass-spectrometry-based assays.

To identify potential drug targets, all druggable proteins on the network were mapped based on known inhibitors cataloged in the Drug Gene Interaction Database (Figure 7F; Table S7D) (Cotto et al., 2018). A small subnetwork of 65 upregulated proteins (median $\log_2 T/N > 0.5$) was identified that includes several matrix metalloproteinases (MMP2, MMP11, MMP12, and MMP14, Figure 7F; Table S7E), suggesting significant regulation of the matrisome. Focusing on the most regulated matrisome proteins (Naba et al., 2012), basement membranes and ECM-affiliated proteins were found mostly downregulated, whereas ECM regulators, including MMPs and Cathepsins, were upregulated (Figure 7G). These findings likely reflect the modulation of TME, with MMPs functioning as key players (Naba et al., 2012). MMP7, MMP11, and MMP12 showed the most significant upregulation in the late and late-like classes in subtype 1 (Figure 7H). Overexpression of MMPs to regulate lung malignancies and their use as therapeutic targets has been reported (Merchant et al., 2017). To evaluate their potential also as biomarker candidates, immunohistochemistry staining for selected MMPs was performed in the tumors of another independent retrospective cohort of 117 early stage patients with survival data. The results showed that strong expression of MMP11 and MMP7 significantly associated with poor overall survival (Figures 7I and S7F; Table S7F), suggesting their potential prognostic value. MMP11 has been previously found overexpressed in NSCLC (Kettunen et al., 2004) and recently reported as a key lung cancer-promoting gene (Yang et al., 2019); however, its protein-level regulation during progression has never been explored. Taken together, we propose MMP11 as a candidate with potential as a biomarker for early detection and treatment of NSCLC for further validation in a larger cohort.

DISCUSSION

Epidemiological studies showed that lung cancer in East Asia is a distinct disease characterized by high prevalence of never-smokers, especially among females, with significantly more frequent *EGFR* activating mutations compared to Caucasian cohorts. Although advances have been made in targeted therapy and immunotherapy for patients in the late stage, early detection and prevention may substantially improve the clinical outcome with economic benefits for patients. However, there are many challenging unmet clinical needs in early stage NSCLC, including relapse after surgical resection for ~20% stage I patients worldwide and poor overall survival (Sawabata et al., 2010). Moreover, patients with different *EGFR* mutations (L858R and Del19) experience different treatment outcomes, different tendencies for malignant pleural effusion, and cancer metastasis. Thus, there is an urgent need to uncover the early processes and progression of oncogenesis, dysregulated molecular pathways, and contributing factors underpinning the distinct features of NSCLC patients in East Asia, which could suggest prognostic biomarkers and novel drug targets. Here, we present the first proteogenomic landscape of East Asian LUAD by deep profiling of tumor and adjacent normal tissues. Our analysis provides a comprehensive insight into the multilayer molecular architecture of early stage LUAD in never-smokers, encompassing somatic mutations, endogenous and exogenous environmental mutational signatures, proteomic subtypes, phosphorylation alterations, and protein co-variation networks capturing functional associations.

The genomic landscape revealed significant differences between our cohort and previous studies comprising mostly smokers, especially in the mutational profiles of known cancer

genes. Even among non-smokers, significantly different mutation frequencies of top-ranking genes indicate different driver alterations between TW and TCGA LUAD cohorts. A notable example is the *RBM10* gene, a LUAD tumor suppressor (Hernandez et al., 2016; Zhao et al., 2017) reported more frequently mutated in males (Cancer Genome Atlas Research, 2014), yet it was found more frequently mutated in the older females of our cohort. RBM10 was observed in a co-regulation network enriched for RNA splicing functions, suggesting that significant loss of RBM10 protein due to mutations could impact its interactions, leading to impaired RNA splicing. The proteogenomic relationships further revealed how genomic features orchestrate the proteome and phosphoproteome in LUAD. The positive correlations of *TP53* mutations with cell cycle genes and phosphoproteins involved in DNA topology regulation and DNA damage response are consistent with known synthetic lethal interactions (Yeo et al., 2016). Co-regulated phosphorylation of MAPK pathway proteins was observed and distinguished patients with high activation associated with *EGFR* and *KRAS* mutations, while low activation coincided with *TP53* mutations, especially in later stages. These findings may provide insight on the role of *TP53* in regulating key pathways in NSCLC.

In this non-smoking cohort, 5 mutational profiles resembling those of endogenous mutagens and environmental agents were identified with distinct age- and gender-related attributes, which was not observed in the TCGA cohort. The high proportion of C>T transitions was linked to the APOBEC mutational signature, which is significantly enriched in females at younger age or with *EGFR*-WT. Recent work reported that APOBEC mutational signatures are the major factors driving early-onset squamous cell carcinomas (Cho et al., 2018). Our results suggest that APOBEC mutagenesis may be a key factor contributing to the early onset of LUAD in females. Additionally, our data revealed increased protein and phosphorylation abundance of APOBEC enzymes and DNA repair pathways in females with high-APOBEC signature, highlighting opportunities for targeted therapeutics. Furthermore, we provide preliminary indication for the potential contribution of environmental agents in the age- and gender-dependent mutagenesis mechanisms, in line with the epidemiological studies on their cancer risk and regional exposure (Loh et al., 2011; Yan et al., 2019). Compared to the TCGA cohort, Nitrosamine and Nitro-PAHs signatures were prominent only in the TW LUAD. In addition to positive correlation with APOBEC signature, the Nitrosamine signature was significantly enriched in stage IB, *EGFR* mutations, female, older age, and non-smokers (Table S2E). The Nitro-PAHs signature was higher in male, older age, and ex- and non-smokers, and inversely correlated with APOBEC signature. The mixed group negatively correlated with APOBEC signature, while positively associated with *EGFR* mutations, female, older age, and non-smoker. These results suggest that different mutational processes contribute to the LUAD in never-smoker East Asian patients in age-, gender- and *EGFR*-mutation-dependent manner. Further exploration in a larger patient cohort and body fluids may offer complementary approaches to early detection and prevention. Nevertheless, mutational signatures are dependent on carcinogens of regional relevance and can be tissue and cell type dependent, which may affect the similarity of associated carcinogen signature. Further studies are warranted to investigate the contribution of environmental carcinogens in promoting tumorigenesis in different regions and populations.

Our findings provide insight for development of new therapeutic strategies for NSCLC in East Asia. The association between higher APOBEC mutational signature and prolonged progression-free survival on the immunotherapy cohort (Hellmann et al., 2018) suggests the potential utility of APOBEC signature to identify patients who may respond to immunotherapy, particularly young female patients with *EGFR*-WT tumors. The main goal of genomic diagnostic profiling is to elucidate driver mutations that predict therapeutic efficacy. Beyond genetic testing, proteome subtyping reveals the molecular heterogeneity of same-stage tumors, differentiating aggressive clinical features among patients in clinically defined early stage. Interestingly, we found that although proteomics subtypes largely recapitulated pathological staging, key genomic features could alter the proteomic taxonomy of specific tumors. The striking finding of significant enrichment of *EGFR*-L858R in stage IB late-like subtype prompted further validation in another retrospectively collected cohort, which confirmed that the two main EGFR activating mutations, L858R and Del19, were associated with the different outcome of EGFR-L858R/Del19 mutations for stage IB patients. Our proteomics-based classification resolved the heterogeneity in *EGFR* mutations and enhanced patient stratification. For Stage IB NSCLC, the effect of adjuvant chemotherapy is still equivocal and only considered for patients with high-risk factors, including poorly differentiated tumors, vascular invasion, tumor >4 cm, and visceral pleural involvement. Based on our findings, *EGFR* mutation testing at stage IB might be helpful for patient stratification to distinguish the high-risk stage IB patients for closer follow-up and possibly adjuvant therapy.

Lastly, *de novo* construction of protein networks not only provided a holistic view of the biological features determining each molecular subtype, but also rationalized the selection of potential drug targets and biomarkers. The integrated network revealed the presence of several immune cell types co-existing in tumor samples with heterogeneous immune cell abundances and a significant role of matrix metalloproteinases in lung cancer progression. Using an independent retrospective cohort, we showed that this druggable protein class is associated with disease outcome. Overall, this study revealed the molecular architecture and hallmarks of tumor progression in early stage LUAD and may enable a path of precision medicine to manage non-smoking lung cancer in East Asia.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead Contact
  - Materials Availability
  - Data and Code Availability

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cell.2020.06.012.

ACKNOWLEDGMENTS

This research was supported by the Next-Generation Pathway of Taiwan Cancer Precision Medicine Program (AS-KPQ-107-TCPMP), the Taiwan Protein Project (AS-KPQ-105-TPP) at Academia Sinica, Ministry of Science and Technology (MOST-106-3114-Y-043F-014, MOST108-2319-B-002-001), and National Taiwan University (NTU-CDP-106R7891) in Taiwan. We thank Drs. Fu-Tong Liu, Karin D. Rodland, and Tzu-Ching Meng for advice on the project and Drs. Chia-Feng Tsai, Chuan-Chih Hsu, and Yasushi Ishihama for comments on phosphoproteomic workflow. We thank the Academia Sinica Common Mass Spectrometry Facilities (AS-CFII-108-107) at the Institute of Chemistry and Institute of Biological Chemistry for mass spectrometry analysis, National Core Facility for Biopharmaceuticals Pharmacogenomics Lab TR6 (MOST108-2319-B-002-001) for NGS technical support, National Center for High-Performance Computing of National Applied Research Laboratories in Taiwan for providing computational and storage resources (MOST108-2319-B-492-001), and Mathematics in Biology Group at the Institute of Statistical Science, Academia Sinica for bioinformatics work. P.-C.Y. acknowledges support from Ministry of Science and Technology (MOST 107-0210-01-19-01, MOST 108-3114-Y-001-002). S.-L.Y. acknowledges support of the Center of Precision Medicine from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE), Taiwan. T.I.R. and J.S.C. are funded by CRUK Centre (C309/A25144), UK. We thank Dr. Serena Nik Zainal for comments on data analysis. We thank Dr. Anguraj Sadanandam and Dr. Gift Nyamundanda for discussions about data analysis. This work was done under the auspices of a memorandum of understanding between the Academia Sinica and the U.S. National Cancer Institute's International Cancer Proteogenome Consortium (ICPC). ICPC encourages international cooperation among institutions and nations in proteogenomic cancer research in which proteogenomic datasets are made available to the public. This work was also done in collaboration with the U.S. National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC).

AUTHOR CONTRIBUTIONS

Conceptualization, T.I.R., J.-S.C., S.-L.Y., J.S.C., H.-Y.C., P.-C.Y., and Yu-Ju Chen; Methodology, Yi-Ju Chen, Y.-H.C., M.-H.L., P.-Y.L., Y.-S.C., P.-S.W., C.-T.L., S.-H.W., K.-Y.S., S.-L.Y., H.-Y.C., and Yu-Ju Chen; Software, Y.-H.C., C.-T.C., T.-C.Y., J.-H.W., P.-Y.T., and H.-Y.C.; Validation, G.-C.C., Y.-L.C., C.-T.W., and H.-Y.C.; Formal Analysis, Yi-Ju Chen, T.I.R., Y.-H.C., C.-T.C., C.-L.H., M.-H.L., Y.-R.C., I.J., F.Z.G., T.-C.Y., J.-H.W., L.R., C.-Y.L., Wei-Hung Chang, P.-Y.T., T.-Y.S., H.-Y.C., and Yu-Ju Chen; Investigation, Yi-Ju Chen, T.I.R., Y.-H.C., M.-H.L., Z.-S.L., K.-T.L., C.-W.C., P.Y.S., C.-T.H., K.-C.H., H.-C.Y., P.-Y.L., Y.-W.L., L.R., Wen-Hsin Chang, Y.-J.H., C.-Y.L., P.-S.W., C.-T.L., J.S.C., and H.-Y.C.; Computation & Statistical Analysis, Yi-Ju Chen, T.I.R., Y.-H.C., C.-T.C., M.-H.L., I.J., F.Z.G., T.-C.Y., J.-H.W., T.-Y.S., and H.-Y.C.; Resources, M.-W.L., M.-S.H., Y.-T.W., T.-Y.S., J.-S.C., S.-L.Y., and Yu-Ju Chen; Data Curation, Yi-Ju Chen, Y.-H.C., C.-L.H., M.-H.L., Y.-T.W., C.-T.L., H.-Y.C., and Yu-Ju Chen; Writing – Original Draft, Yi-Ju Chen, T.I.R., Y.-H.C., C.-L.H., M.-H.L., G.-C.C., Y.-L.C., C.-T.W., M.-W.L., K.-T.L., S.-L.Y., J.S.C., H.-Y.C., and Yu-Ju Chen; Writing – Review & Editing, Yi-Ju Chen, T.I.R., C.-L.H., H.-W.C., I.J., F.Z.G., A.I.R., H.R., J.S.C., H.-Y.C., P.-C.Y., and Yu-Ju Chen; Visualization, Yi-Ju Chen, T.I.R., Y.-H.C., C.-T.C., C.-L.H., M.-H.L., I.J., F.Z.G., K.-T.L., L.R., C.-Y.L., H.-Y.C., P.-C.Y., and Yu-Ju Chen; Supervision, C.-L.H., H.-W.C., Y.-R.C., I.J., T.-Y.S., J.-S.C., S.-L.Y., J.S.C., H.-Y.C., P.-C.Y., and Yu-Ju Chen; Project Administration, Yi-Ju Chen, Y.-H.C., C.-T.C., C.-L.H., H.R., S.-L.Y., J.S.C., H.-Y.C., and Yu-Ju Chen; Funding Acquisition, S.-L.Y., J.S.C., H.-Y.C., and Yu-Ju Chen.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 15, 2019
Revised: March 13, 2020
Accepted: June 3, 2020
Published: July 9, 2020

REFERENCES

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Borresen-Dale, A.L., et al. (2013a). Signatures of mutational processes in human cancer. Nature *500*, 415–421.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J., and Stratton, M.R. (2013b). Deciphering signatures of mutational processes operative in human cancer. Cell Rep. *3*, 246–259.

Amin, M.B.; American Joint Committee on Cancer, and American Cancer Society (2017). AJCC cancer staging manual, Eight edition, M.B. Amin, S.B. Edge, D.M. Gress, and L.R. Meyer, eds. (Chicago IL: American Joint Committee on Cancer, Springer).

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Andreatta, M., and Nielsen, M. (2016). Gapped sequence alignment using artificial neural networks: application to the MHC class I system. Bioinformatics *32*, 511–517.

Bach, P.B. (2009). Smoking as a Factor in Causing Lung Cancer. JAMA *301*, 539–541.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. *68*, 394–424.

Brunet, J.P., Tamayo, P., Golub, T.R., and Mesirov, J.P. (2004). Metagenes and molecular pattern discovery using matrix factorization. Proc. Natl. Acad. Sci. USA *101*, 4164–4169.

Campbell, J.D., Alexandrov, A., Kim, J., Wala, J., Berger, A.H., Pedamallu, C.S., Shukla, S.A., Guo, G., Brooks, A.N., Murray, B.A., et al. (2016). Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. Nat. Genet. *48*, 607–616.

Cancer Genome Atlas Research Network (2014). Comprehensive molecular profiling of lung adenocarcinoma. Nature *511*, 543–550.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBio. cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. *2*, 401–404.

Chen, L., Ge, B., Casale, F.P., Vasquez, L., Kwan, T., Garrido-Martín, D., Watt, S., Yan, Y., Kundu, K., Ecker, S., et al. (2016). Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell *167*, 1398–1414.

Cho, R.J., Alexandrov, L.B., den Breems, N.Y., Atanasova, V.S., Farshchian, M., Purdom, E., Nguyen, T.N., Coarfa, C., Rajapakshe, K., Prisco, M., et al.

(2018). APOBEC mutation drives early-onset squamous cell carcinomas in recessive dystrophic epidermolysis bullosa. Sci. Transl. Med. *10* https://doi.org/10.1126/scitranslmed.aas9668.

Corsello, S.M., Nagari, R.T., Spangler, R.D., Rossen, J., Kocak, M., Bryan, J.G., Humeidi, R., Peck, D., Wu, X., Tang, A.A., et al. (2019). Non-oncology drugs are a source of previously unappreciated anti-cancer activity. bioRxiv. https://doi.org/10.1038/s43018-019-0018-6.

Cotto, K.C., Wagner, A.H., Feng, Y.Y., Kiwala, S., Coffman, A.C., Spies, G., Wollam, A., Spies, N.C., Griffith, O.L., and Griffith, M. (2018). DGIdb 3.0: a redesign and expansion of the drug-gene interaction database. Nucleic Acids Res. *46*, D1068–D1073.

Cox, J., and Mann, M. (2012). 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. BMC Bioinformatics *13* (*Suppl 16*), S12.

Dimayacyac-Esleta, B.R., Tsai, C.F., Kitata, R.B., Lin, P.Y., Choong, W.K., Lin, T.D., Wang, Y.T., Weng, S.H., Yang, P.C., Arco, S.D., et al. (2015). Rapid High-pH Reverse Phase StageTip for Sensitive Small-Scale Membrane Proteomic Profiling. Anal. Chem. *87*, 12016–12023.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dorfer, V., Pichler, P., Stranzl, T., Stadlmann, J., Taus, T., Winkler, S., and Mechtler, K. (2014). MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. J. Proteome Res. *13*, 3679–3684.

Du, Z., and Lovly, C.M. (2018). Mechanisms of receptor tyrosine kinase activation in cancer. Mol. Cancer *17*, 58.

Eng, J.K., McCormack, A.L., and Yates, J.R., III. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. *5*, 976–989.

Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. (2013). Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. Sci. Signal. *6*, pl1.

Gao, Q., Zhu, H., Dong, L., Shi, W., Chen, R., Song, Z., Huang, C., Li, J., Dong, X., Zhou, Y., et al. (2019). Integrated Proteogenomic Characterization of HBV-Related Hepatocellular Carcinoma. Cell *179*, 561–577.e22.

Ghandi, M., Huang, F.W., Jane-Valbuena, J., Kryukov, G.V., Lo, C.C., McDonald, E.R., 3rd, Barretina, J., Gelfand, E.T., Bielski, C.M., Li, H., et al. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature *569*, 503–508.

Gillette, M.A., Satpathy, S., Cao, S., Dhanasekaran, S.M., Vasaikar, S.V., Krug, K., Petralia, F., Li, Y., Liang, W.-W., Reva, B., et al. (2020). Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. Cell *183*, this issue, 200–225.

Gushgari, A.J., and Halden, R.U. (2018). Critical review of major sources of human exposure to N-nitrosamines. Chemosphere *210*, 1124–1136.

Hellmann, M.D., Nathanson, T., Rizvi, H., Creelan, B.C., Sanchez-Vega, F., Ahuja, A., Ni, A., Novik, J.B., Mangarin, L.M.B., Abu-Akeel, M., et al. (2018). Genomic Features of Response to Combination Immunotherapy in Patients with Advanced Non-Small-Cell Lung Cancer. Cancer Cell *33*, 843–852.e4.

Hernandez, J., Bechara, E., Schlesinger, D., Delgado, J., Serrano, L., and Valcarcel, J. (2016). Tumor suppressor properties of the splicing regulatory factor RBM10. RNA Biol. *13*, 466–472.

Hua, X., Hyland, P.L., Huang, J., Song, L., Zhu, B., Caporaso, N.E., Landi, M.T., Chatterjee, N., and Shi, J. (2016). MEGSA: A Powerful and Flexible Framework for Analyzing Mutual Exclusivity of Tumor Mutations. Am. J. Hum. Genet. *98*, 442–455.

Imielinski, M., Berger, A.H., Hammerman, P.S., Hernandez, B., Pugh, T.J., Hodis, E., Cho, J., Suh, J., Capelletti, M., Sivachenko, A., et al. (2012). Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell *150*, 1107–1120.

Jakszyn, P., and Gonzalez, C.A. (2006). Nitrosamine and related food intake and gastric and oesophageal cancer risk: a systematic review of the epidemiological evidence. World J. Gastroenterol. *12*, 4296–4303.

Jemal, A., Miller, K.D., Ma, J., Siegel, R.L., Fedewa, S.A., Islami, F., Devesa, S.S., and Thun, M.J. (2018). Higher Lung Cancer Incidence in Young Women Than Young Men in the United States. N. Engl. J. Med. *378*, 1999–2009.

Jiang, T., Shi, T., Zhang, H., Hu, J., Song, Y., Wei, J., Ren, S., and Zhou, C. (2019). Tumor neoantigens: from basic research to clinical applications. J. Hematol. Oncol. *12*, 93.

Kawaguchi, T., Matsumura, A., Fukai, S., Tamura, A., Saito, R., Zell, J.A., Maruyama, Y., Ziogas, A., Kawahara, M., and Ignatius Ou, S.H. (2010). Japanese ethnicity compared with Caucasian ethnicity and never-smoking status are independent favorable prognostic factors for overall survival in non-small cell lung cancer: a collaborative epidemiologic study of the National Hospital Organization Study Group for Lung Cancer (NHSGLC) in Japan and a Southern California Regional Cancer Registry databases. J. Thorac. Oncol. *5*, 1001–1010.

Kelly, W.J., Shah, N.J., and Subramaniam, D.S. (2018). Management of Brain Metastases in Epidermal Growth Factor Receptor Mutant Non-Small-Cell Lung Cancer. Front. Oncol. *8*, 208.

Kettunen, E., Anttila, S., Seppanen, J.K., Karjalainen, A., Edgren, H., Lindstrom, I., Salovaara, R., Nissen, A.M., Salo, J., Mattson, K., et al. (2004). Differentially expressed genes in nonsmall cell lung cancer: expression profiling of cancer-related genes in squamous cell lung cancer. Cancer Genet. Cytogenet. *149*, 98–106.

Kim, H., Chen, J., and Yu, X. (2007). Ubiquitin-binding protein RAP80 mediates BRCA1-dependent DNA damage response. Science *316*, 1202–1205.

Kim, M.S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., et al. (2014). A draft map of the human proteome. Nature *509*, 575–581.

Koch, H., Wilhelm, M., Ruprecht, B., Beck, S., Frejno, M., Klaeger, S., and Kuster, B. (2016). Phosphoproteome Profiling Reveals Molecular Mechanisms of Growth-Factor-Mediated Kinase Inhibitor Resistance in EGFR-Overexpressing Cancer Cells. J. Proteome Res. *15*, 4490–4504.

Kucab, J.E., Zou, X., Morganella, S., Joel, M., Nanda, A.S., Nagy, E., Gomez, C., Degasperi, A., Harris, R., Jackson, S.P., et al. (2019). A Compendium of Mutational Signatures of Environmental Agents. Cell *177*, 821–836.e16.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559.

Lapek, J.D., Jr., Greninger, P., Morris, R., Amzallag, A., Pruteanu-Malinici, I., Benes, C.H., and Haas, W. (2017). Detection of dysregulated protein-association networks by high-throughput proteomics predicts cancer vulnerabilities. Nat. Biotechnol. *35*, 983–989.

Lee, D.D., and Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature *401*, 788–791.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760.

Lin, Z.P., Zhu, Y.L., and Ratner, E.S. (2018). Targeting Cyclin-Dependent Kinases for Treatment of Gynecologic Cancers. Front. Oncol. *8*, 303.

Liu, S., Shiotani, B., Lahiri, M., Marechal, A., Tse, A., Leung, C.C., Glover, J.N., Yang, X.H., and Zou, L. (2011). ATR autophosphorylation as a molecular switch for checkpoint activation. Mol. Cell *43*, 192–202.

Loh, Y.H., Jakszyn, P., Luben, R.N., Mulligan, A.A., Mitrou, P.N., and Khaw, K.T. (2011). N-Nitroso compounds and cancer incidence: the European Prospective Investigation into Cancer and Nutrition (EPIC)-Norfolk Study. Am. J. Clin. Nutr. *93*, 1053–1061.

Luo, W., Tian, P., Wang, Y., Xu, H., Chen, L., Tang, C., Shu, Y., Zhang, S., Wang, Z., Zhang, J., et al. (2018). Characteristics of genomic alterations of lung adenocarcinoma in young never-smokers. Int. J. Cancer *143*, 1696–1705.

Maslyk, M., Janeczko, M., Martyna, A., and Kubinski, K. (2017). CX-4945: the protein kinase CK2 inhibitor and anti-cancer drug shows anti-fungal activity. Mol. Cell. Biochem. *435*, 193–196.

Mayakonda, A., Lin, D.C., Assenov, Y., Plass, C., and Koeffler, H.P. (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. *28*, 1747–1756.

Merchant, N., Nagaraju, G.P., Rajitha, B., Lammata, S., Jella, K.K., Buchwald, Z.S., Lakka, S.S., and Ali, A.N. (2017). Matrix metalloproteinases: their functional role in lung cancer. Carcinogenesis *38*, 766–780.

Mhanna, L., Guibert, N., Milia, J., and Mazieres, J. (2019). When to Consider Immune Checkpoint Inhibitors in Oncogene-Driven Non-Small Cell Lung Cancer? Curr. Treat. Options Oncol. *20*, 60.

Moorthy, B., Chu, C., and Carlin, D.J. (2015). Polycyclic aromatic hydrocarbons: from metabolism to lung cancer. Toxicol. Sci. *145*, 5–15.

Mross, K., Richly, H., Frost, A., Scharr, D., Nokay, B., Graeser, R., Lee, C., Hilbert, J., Goeldner, R.G., Fietz, O., et al. (2016). A phase I study of BI 811283, an Aurora B kinase inhibitor, in patients with advanced solid tumors. Cancer Chemother. Pharmacol. *78*, 405–417.

Naba, A., Clauser, K.R., Hoersch, S., Liu, H., Carr, S.A., and Hynes, R.O. (2012). The matrisome: in silico definition and in vivo characterization by proteomics of normal and tumor extracellular matrices. Mol Cell Proteomics *11*. https://doi.org/10.1074/mcp.M111.014647.

Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods *12*, 453–457.

Qiu, W.G., Polotskaia, A., Xiao, G., Di, L., Zhao, Y., Hu, W., Philip, J., Hendrickson, R.C., and Bargonetti, J. (2017). Identification, validation, and targeting of the mutant p53-PARP-MCM chromatin axis in triple negative breast cancer. NPJ Breast Cancer *3*. https://doi.org/10.1038/s41523-016-0001-7.

Roberts, S.A., Lawrence, M.S., Klimczak, L.J., Grimm, S.A., Fargo, D., Stojanov, P., Kiezun, A., Kryukov, G.V., Carter, S.L., Saksena, G., et al. (2013). An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nat. Genet. *45*, 970–976.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. *11*, R25.

Roper, N., Gao, S., Maity, T.K., Banday, A.R., Zhang, X., Venugopalan, A., Cultraro, C.M., Patidar, R., Sindiri, S., Brown, A.L., et al. (2019). APOBEC Mutagenesis and Copy-Number Alterations Are Drivers of Proteogenomic Tumor Evolution and Heterogeneity in Metastatic Thoracic Tumors. Cell Rep *26*, 2651–2666.e6.

Roumeliotis, T.I., Williams, S.P., Goncalves, E., Alsinet, C., Del Castillo Velasco-Herrera, M., Aben, N., Ghavidel, F.Z., Michaut, M., Schubert, M., Price, S., et al. (2017). Genomic Determinants of Protein Abundance Variation in Colorectal Cancer Cells. Cell Rep. *20*, 2201–2214.

Ryan, C.J., Kennedy, S., Bajrami, I., Matallanas, D., and Lord, C.J. (2017). A Compendium of Co-regulated Protein Complexes in Breast Cancer Reveals Collateral Loss Events. Cell Syst *5*, 399–409.e5.

Samet, J.M., Avila-Tang, E., Boffetta, P., Hannan, L.M., Olivo-Marston, S., Thun, M.J., and Rudin, C.M. (2009). Lung cancer in never smokers: clinical epidemiology and environmental risk factors. Clin. Cancer Res. *15*, 5626–5645.

Sawabata, N., Asamura, H., Goya, T., Mori, M., Nakanishi, Y., Eguchi, K., Koshiishi, Y., Okumura, M., Miyaoka, E., Fujii, Y., et al. (2010). Japanese Lung Cancer Registry Study: first prospective enrollment of a large number of surgical and nonsurgical cases in 2002. J. Thorac. Oncol. *5*, 1369–1375.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.

Shi, Y., Au, J.S., Thongprasert, S., Srinivasan, S., Tsai, C.M., Khoa, M.T., Heeroma, K., Itoh, Y., Cornelio, G., and Yang, P.C. (2014). A prospective, molecular epidemiology study of EGFR mutations in Asian patients with advanced non-small-cell lung cancer of adenocarcinoma histology (PIONEER). J. Thorac. Oncol. *9*, 154–162.

Shukla, S.A., Rooney, M.S., Rajasagi, M., Tiao, G., Dixon, P.M., Lawrence, M.S., Stevens, J., Lane, W.J., Dellagatta, J.L., Steelman, S., et al. (2015). Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. Nat. Biotechnol. *33*, 1152–1158.

Su, K.Y., Chen, H.Y., Li, K.C., Kuo, M.L., Yang, J.C., Chan, W.K., Ho, B.C., Chang, G.C., Shih, J.Y., Yu, S.L., et al. (2012). Pretreatment epidermal growth factor receptor (EGFR) T790M mutation predicts shorter EGFR tyrosine kinase inhibitor response duration in patients with non-small-cell lung cancer. J. Clin. Oncol. *30*, 433–440.

Suda, K., Tomizawa, K., and Mitsudomi, T. (2010). Biological and clinical significance of KRAS mutations in lung cancer: an oncogenic driver that contrasts with EGFR mutation. Cancer Metastasis Rev. *29*, 49–60.

Sun, S., Schiller, J.H., and Gazdar, A.F. (2007). Lung cancer in never smokers–a different disease. Nat. Rev. Cancer *7*, 778–790.

Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. *47*, D607–D613.

Tam, I.Y., Leung, E.L., Tin, V.P., Chua, D.T., Sihoe, A.D., Cheng, L.C., Chung, L.P., and Wong, M.P. (2009). Double EGFR mutants containing rare EGFR mutant types show reduced in vitro response to gefitinib compared with common activating missense mutations. Mol. Cancer Ther. *8*, 2142–2151.

Taus, T., Köcher, T., Pichler, P., Paschke, C., Schmidt, A., Henrich, C., and Mechtler, K. (2011). Universal and confident phosphorylation site localization using phosphoRS. J. Proteome Res. *10*, 5354–5362.

Tomasello, C., Baldessari, C., Napolitano, M., Orsi, G., Grizzi, G., Bertolini, F., Barbieri, F., and Cascinu, S. (2018). Resistance to EGFR inhibitors in non-small cell lung cancer: Clinical management and future perspectives. Crit. Rev. Oncol. Hematol. *123*, 149–161.

Tomkinson, A.E., Howes, T.R., and Wiest, N.E. (2013). DNA ligases as therapeutic targets. Transl. Cancer Res. *2*, 1219.

Tsai, M.J., Hung, J.Y., Lee, M.H., Kuo, C.Y., Tsai, Y.C., Tsai, Y.M., Liu, T.C., Yang, C.J., Huang, M.S., and Chong, I.W. (2018). Better Progression-Free Survival in Elderly Patients with Stage IV Lung Adenocarcinoma Harboring Uncommon Epidermal Growth Factor Receptor Mutations Treated with the First-line Tyrosine Kinase Inhibitors. Cancers (Basel) *10*. https://doi.org/10.3390/cancers10110434.

Tseng, C.H., Tsuang, B.J., Chiang, C.J., Ku, K.C., Tseng, J.S., Yang, T.Y., Hsu, K.H., Chen, K.C., Yu, S.L., Lee, W.C., et al. (2019). The Relationship Between Air Pollution and Lung Cancer in Nonsmokers in Taiwan. J. Thorac. Oncol. *14*, 784–792.

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M.Y., Geiger, T., Mann, M., and Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat. Methods *13*, 731–740.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics *43*, 11.10.1–11.10.33.

Vasaikar, S., Huang, C., Wang, X., Petyuk, V.A., Savage, S.R., Wen, B., Dou, Y., Zhang, Y., Shi, Z., Arshad, O.A., et al. (2019). Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. Cell *177*, 1035–1049.e19.

Wang, B., Matsuoka, S., Ballif, B.A., Zhang, D., Smogorzewska, A., Gygi, S.P., and Elledge, S.J. (2007). Abraxas and RAP80 form a BRCA1 protein complex required for the DNA damage response. Science *316*, 1194–1198.

Wang, C., Yin, R., Dai, J., Gu, Y., Cui, S., Ma, H., Zhang, Z., Huang, J., Qin, N., Jiang, T., et al. (2018a). Whole-genome sequencing reveals genomic signatures associated with the inflammatory microenvironments in Chinese NSCLC patients. Nat. Commun. *9*, 2054.

Wang, H.Z., Yang, S.H., Li, G.Y., and Cao, X. (2018b). Subunits of human condensins are potential therapeutic targets for cancers. Cell Div. *13*, 2.

Wang, X., and Zhang, B. (2013). customProDB: an R package to generate customized protein databases from RNA-Seq data. Bioinformatics *29*, 3235–3237.

Weidner, C., Fischer, C., and Sauer, S. (2014). PHOXTRACK-a tool for interpreting comprehensive datasets of post-translational modifications of proteins. Bioinformatics *30*, 3410–3411.

Wilkerson, M.D., and Hayes, D.N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics *26*, 1572–1573.

Wu, Y.L., Liam, C.K., Zhou, C.C., Wu, G., Liu, X.Q., Zhong, Z.Y., Lu, S., Cheng, Y., Han, B.H., Chen, L., et al. (2013). First-Line Erlotinib Versus Cisplatin/Gemcitabine (Gp) in Patients with Advanced Egfr Mutation-Positive Non-Small-Cell Lung Cancer (Nsclc): Interim Analyses from the Phase 3, Open-Label, Ensure Study. J. Thorac. Oncol. *8*, S603–S604.

Yan, D., Wu, S., Zhou, S., Tong, G., Li, F., Wang, Y., and Li, B. (2019). Characteristics, sources and health risk assessment of airborne particulate PAHs in Chinese cities: A review. Environ. Pollut. *248*, 804–814.

Yang, H., Jiang, P., Liu, D., Wang, H.Q., Deng, Q., Niu, X., Lu, L., Dai, H., Wang, H., and Yang, W. (2019). Matrix Metalloproteinase 11 Is a Potential Therapeutic Target in Lung Adenocarcinoma. Mol. Ther. Oncolytics *14*, 82–93.

Yang, C.J., Yang, J.C.H., and Yang, P.C. (2020). Precision management of advanced non-small cell lung cancer. Annu. Rev. Med. *2020*, 117–136.

Yeo, C.Q.X., Alexander, I., Lin, Z., Lim, S., Aning, O.A., Kumar, R., Sangthongpitag, K., Pendharkar, V., Ho, V.H.B., and Cheok, C.F. (2016). p53 Maintains Genomic Stability by Preventing Interference between Transcription and Replication. Cell Rep. *15*, 132–146.

Yoshihara, K., Shahmoradgoli, M., Martinez, E., Vegesna, R., Kim, H., Torres-Garcia, W., Trevino, V., Shen, H., Laird, P.W., Levine, D.A., et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. Nat. Commun. *4*, 2612.

Zhang, X., Belkina, N., Jacob, H.K., Maity, T., Biswas, R., Venugopalan, A., Shaw, P.G., Kim, M.S., Chaerkady, R., Pandey, A., et al. (2015). Identifying novel targets of oncogenic EGF receptor signaling in lung cancer through global phosphoproteomics. Proteomics *15*, 340–355.

Zhang, H., Liu, T., Zhang, Z., Payne, S.H., Zhang, B., McDermott, J.E., Zhou, J.Y., Petyuk, V.A., Chen, L., Ray, D., et al. (2016). Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian. Cancer Cell *166*, 755–765.

Zhang, X.C., Wang, J., Shao, G.G., Wang, Q., Qu, X., Wang, B., Moy, C., Fan, Y., Albertyn, Z., Huang, X., et al. (2019). Comprehensive genomic and immunological characterization of Chinese non-small cell lung cancer patients. Nat. Commun. *10*, 1772.

Zhao, J., Sun, Y., Huang, Y., Song, F., Huang, Z., Bao, Y., Zuo, J., Saffen, D., Shao, Z., Liu, W., et al. (2017). Functional analysis reveals that RBM10 mutations contribute to lung adenocarcinoma pathogenesis by deregulating splicing. Sci. Rep. *7*, 40488.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
MMP-2 antibody	Santa Cruz Biotechnology, Inc.	Cat# sc-53630; RRID: AB_784594
MMP-7 antibody	Santa Cruz Biotechnology, Inc.	Cat# sc-80205; RRID: AB_1126314
MMP-11 antibody	Santa Cruz Biotechnology, Inc.	Cat# sc-58381; RRID: AB_2144725
Biological Samples		
Tumor and adjacent normal tissues	National Taiwan University Hospital	This study
Blood samples	National Taiwan University Hospital and Taichung Veterans General Hospital	This study
Chemicals, Peptides, and Recombinant Proteins		
Ni-NTA silica resin	QIAGEN	Cat No. 31014
Novocastra Epitope Retrieval Solution pH 6	Leica	Cat No. RE7113
Critical Commercial Assays		
AllPrep DNA/RNA/miRNA Universal Kit	QIAGEN	Cat No. 80224
QIAamp DNA Blood Mini Kit	QIAGEN	Cat No. 51106
Qubit dsDNA BR Assay Kit	Thermo Fisher	Cat No. Q32853
Qubit dsDNA HS Assay Kit	Thermo Fisher	Cat No. Q32854
DNeasy PowerClean Cleanup Kit	QIAGEN	Cat No. 12877-50
TruSeq DNA PCR-Free Library Prep Kit	Illumina	Cat No. 20015963
TruSeq Stranded Total RNA Library Prep Human/Mouse/Rat	Illumina	Cat No. 20020597
SureSelect XT2 Library Prep Kit	Agilent	Cat No. 5500-0131
SureSelect XT2 Pre-Capture Box1	Agilent	Cat No. 5190-4076
SureSelect XT2 Human All Exon V6+COSMIC	Agilent	Cat No. 5190-9311
TMT10plex Isobaric Label Reagent Set, 3 × 0.8 mg	Thermo Fisher	Cat No. 90111
UltraVision Quanto Detection System HRP DAB	Thermo Fisher	Cat No. TL-125-QHD
Pierce™ BCA Protein Assay Kit	Thermo Fisher	Catalog: 23225
Deposited Data		
30 validated cancer signatures	COSMIC	https://cancer.sanger.ac.uk/cosmic/signatures_v2
54 carcinogen signatures	Kucab et al., 2019	https://pubchem.ncbi.nlm.nih.gov
TCGA lung adenocarcinoma	Campbell et al., 2016; Cancer Genome Atlas Research, 2014; Imielinski et al., 2012	https://genome.ucsc.edu/
COSMIC onco-driver and suppressor gene data	COSMIC	https://cancer.sanger.ac.uk/census
Mutational Signatures (v2 - March 2015)	COSMIC	https://cancer.sanger.ac.uk/cosmic/signatures_v2
TCGA lung adenocarcinoma with APOBEC signature	TCGA	https://xenabrowser.net/
Software and Algorithms		
FastQC (v0.11.7)		https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Trimmomatic (v0.36)	Bolger et al., 2014	http://www.usadellab.org/cms/?page=trimmomatic

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
BWA (v0.7.17)	Li and Durbin, 2009	http://bio-bwa.sourceforge.net/
GATK 4.0	Van der Auwera et al., 2013	https://software.broadinstitute.org/gatk/
STAR2 (STAR-2.6.0c)	Dobin et al., 2013	https://portal.rc.fas.harvard.edu/apps/modules/centos7/STAR/2.6.0c-fasrc01
htseq-count algorithm	Anders et al., 2015	https://htseq.readthedocs.io/en/release_0.11.1/
edgeR	Robinson et al., 2010	https://bioconductor.org/packages/release/bioc/html/edgeR.html
trimmed mean of M-values (TMM) algorithm	Robinson and Oshlack, 2010	http://bioinformatics.sph.harvard.edu/bcbioRNASeq////////reference/tmm.html
Non-negative matrix (NMF) algorithm	Brunet et al., 2004	https://cran.r-project.org/web/packages/NMF/index.html
cBioPortal	Cerami et al., 2012	https://www.cbioportal.org/
APOBEC enrichment	Roberts et al., 2013	http://www.bioconductor.org/packages/devel/bioc/vignettes/maftools/inst/doc/maftools.html#1_introduction
Proteome Discoverer 2.1	Thermo Fisher	https://www.thermofisher.com/order/catalog/product/OPTON-30945?SID=srch-srp-OPTON-30945
Mascot (version 2.3.2)	Matrix Science	http://www.matrixscience.com/
SEQUEST	Eng et al., 1994	http://proteomicswiki.com/wiki/index.php/SEQUEST
MS Amanda	Dorfer et al., 2014	http://ms.imp.ac.at/?goto=msamanda
ptmRS	Taus et al., 2011	https://omictools.com/ptmrs-tool
psych R package	RDocumentation	https://www.rdocumentation.org/packages/psych/versions/1.8.12
customProDB (R package)	Wang and Zhang, 2013	http://bioconductor.org/packages/release/bioc/html/customProDB.html
SPSS 15.0	IBM	https://www.ibm.com/au-en/products/spss-statistics
GRC37/hg19	NCBI	https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/
SwissProt human protein database (version 2016.05, 20213 entries)	UniProt	https://www.uniprot.org/statistics/Swiss-Prot
PHOXTRACK	Weidner et. al. 2014	http://phoxtrack.molgen.mpg.de/
LIMIX (for QTL)	Chen et al., 2016	https://limix.readthedocs.io/en/stable/qtl.html
ConsensusClusterPlus (R package)	Wilkerson and Hayes, 2010	https://bioconductor.org/packages/release/bioc/html/ConsensusClusterPlus.html
Perseus 1.6	Tyanova et. al. 2016	http://www.perseus-framework.org
WGCNA library (R package)	Langfelder and Horvath, 2008	https://cran.r-project.org/web/packages/WGCNA/index.html
Cytoscape	Shannon et. al. 2003	http://www.cytoscape.org
Morpheus	Broad Institute	https://software.broadinstitute.org/morpheus/
ESTIMATE algorithm	Yoshihara et al., 2013	https://sourceforge.net/projects/estimateproject/
POLYSOLVER	Shukla et al., 2015	https://software.broadinstitute.org/cancer/cga/polysolver
NetMHC (v4.0)	Andreatta and Nielsen, 2016	http://www.cbs.dtu.dk/services/NetMHC/

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Yu-Ju Chen (yujuchen@gate.sinica.edu.tw).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

Genomics and proteomics data reference in this study will be available in publicly accessible NIH-designated data repositories and portals including the database of Genotypes and Phenotypes (dbGaP) (accession number: phs001954.v1.p1) and the NCI Proteomics Data Commons (PDC, accession number: PDC000219 and PDC000220), along with demographic and clinical data. Raw data files and the processed data for proteomic and phosphoproteomic analyses reported in this paper are uploaded to the NCI Proteomics Data Commons and can be accessed at: https://pdc.cancer.gov/pdc/study/PDC000219 and https://pdc.cancer.gov/pdc/study/PDC000220

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Clinical Specimens

There are three clinical cohorts collected for proteogenomics analysis, MMPs clinical validation and *EGFR*-subtype survival analysis. The Research Ethics Committees of Academia Sinica, National Taiwan University Hospital (NTUH) and Taichung Veterans General Hospital (TCVGH) approved the study and all patients provided written informed consent. For proteogenomics analysis, lung cancer and adjacent normal tissues and blood samples were prospectively collected from 108 patients from NTUH between July 2016 and July 2018. Clinical information of individual patients including age, gender, smoking status, histology, stage, EGFR status and primary tumor location is listed in Table S1**A**. The inclusion criteria recruited newly diagnosed, treatment-naive patients undergoing primary surgery for lung adenocarcinoma. For validation of MMP expression by IHC, 134 FFPE specimen tissues were retrospectively collected from NTHU (Stage IA, n = 24; Stage IB, n = 75, Stage II, n = 18, Stage III-IV, n = 17, Table S7**E)**. For comparison of clinical outcome between the *EGFR*-L858R and *EGFR*-Del19 patients at late-like stage, 208 patients were recruited from TCVGH (Stage IA, n = 143; Stage IB, without adjuvant chemotherapy, n = 65). The clinical information regarding patient history, status of surgery along with relevant diagnostic information and survival records were also obtained from TCVGH.

For proteogenomics analysis, after gross examination, non-necrotic tumor and adjacent normal tissues were excised from resected tumor specimens. Only tissues with weight greater than 40 mg were divided into two cyrotubes and stored in liquid nitrogen less than 30 min after resection. Paired adjacent normal tissues were taken from the most distant site relative to the tumor bolder. Finally, the H&E staining and pathology reports were utilized to identify and qualify lung adenocarcinoma cases with at least 60% tumor cell nuclei and less than 20% necrosis for this study. The patients treated with neoadjuvant chemotherapy or molecular targeting therapy were excluded for this study. Clinicopathological parameters, including age, gender, smoking status, tumor site, tumor size, tumor stage, *EGFR* mutation status, and clinical treatment were collected. Pathological staging was based on the American Joint Committee on Cancer 8th edition stage (Amin et al., 2017). The *EGFR* mutations were detected by MALDI-TOF MS performed by ISO15189-certified TR6 Pharmacogenomics Lab (PG Lab), National Core Facility for Biopharmaceuticals (NCFB) of NTU as previously described (Su et al., 2012). The lung cancer TNM (tumor, node, and metastasis) staging was done according to the 7th edition of the American Joint Committee on Cancer (AJCC) staging system.

To compare the clinical outcome of *EGFR* mutant patients, we compare the disease-free survival (DFS) and overall survival (OS) among different *EGFR* mutation subtypes and wild type in complete surgical resected stage IA and IB lung adenocarcinoma patients diagnosed and treated at TCVGH in Taiwan. To be eligible for the validation study, patients were required to have pathological confirmation of lung adenocarcinoma and pathological stage IA and IB disease after surgical resection and had a clear survival follow-up. Patients were excluded if they had advanced disease, active second malignancy, or any condition that may influence the outcome evaluation, such as irregular follow-up, neoadjuvant treatment with chemotherapy or EGFR-TKI. Clinical data included patients' age, gender, smoking status, tumor stage, the Eastern Cooperative Oncology Group performance status (ECOG PS), *EGFR* mutation status, treatment history, and outcome variables.

METHOD DETAILS

Genomic and Transcriptomic Analysis

Whole Exome Sequencing and mRNA Sequencing

Genomic DNA and total RNA were isolated from frozen tumor and adjacent normal tissues using AllPrep DNA/RNA/miRNA Universal Kit (QIAGEN), simultaneously. Genomic DNA in blood buffy coat was isolated using QIAamp DNA Blood Mini Kit (QIAGEN). DNA and RNA quality was confirmed using Bioanalyzer 2100 (Agilent Technologies) and Nanodrop 2000 spectrometer (ThermoFisher). Whole

exome sequencing (WES) were performed using 400 ng of purified genomic DNA by DNeasy PowerClean Cleanup Kit (QIAGEN) from tumor, adjacent normal tissues and buffy coat for library preparation, respectively. SureSelectXT2 target Enrichment System (Agilent Technologies) was used for exome hybrid capture of WES followed by paired-end multiplexed sequencing. The library preparation of mRNA sequencing was performed on 1 μg of total RNA with DNase I treatment using TruSeq Stranded Total RNA Sample Preparation Kit (Illumina). The concentrations of all libraries were quantified by Qubit assays (Thermo Fisher) and the sizes of libraries were measured by Agilent TapeStation (Agilent). The paired-end sequencing was performed using HiSeq 4000 (Illumina). The average mappable depth was 150X for WES, and the averaged mappable reads of RNA sequencing reached 50 million paired reads.

Sequencing Reads Data Preprocessing

Sequencing reads data was checked for quality and adaptor/primer sequence contamination first by FastQC (v0.11.7). Adaptor sequences and unqualified bases were trimmed by Trimmomatic (v0.36) (Bolger et al., 2014). The threshold was average quality < 20 per sliding window with 4-base. After trimming, reads with length < 36 bases were dropped out.

Somatic Variants Detection

Trimmed paired-end reads were further aligned to human reference genome GRC37/hg19 by BWA (v0.7.17) (Li and Durbin, 2009). Because of systematic technical error from sequencing machines, reads' base quality scores were recalibrated by GATK 4.0 (Van der Auwera et al., 2013) and then somatic mutations were detected by MuTect2 (embedded in GATK).

Mutual Exclusivity Analysis of Mutation

Mutually exclusive mutated gene pairs were identified with the MEGSA tool (Hua et al., 2016) in RStudio. For this analysis we used 39 genes with frequency > 10% and 18 genes mapping to KEGG pathways (Frequency > 5%). The settings included: nSimu = 100, nPairStart = 10, maxSize = 3.

Mutation frequency in TW and Previous Studies

Mutation frequencies for three previous lung adenocarcinoma studies (Campbell et al., 2016; Cancer Genome Atlas Research, 2014; Imielinski et al., 2012) were downloaded from the cBioPortal (Cerami et al., 2012; Gao et al., 2013). The frequencies of all genes were compared with those from the TW cohort using Spearman's correlation similarity matrix. Note that 58% of the cases included in Campbell et al. were part of the Imielinski et al. and The Cancer Genome Atlas studies.

RNA Gene Expression

Trimmed paired-end reads were also mapped to human reference genome GRC37/hg19 by STAR2 (STAR-2.6.0c) (Dobin et al., 2013). Ensemble GRC37 gene annotation and the htseq-count algorithm (Anders et al., 2015) were used to quantify gene expression by raw read counts. In order to adjust different sequencing throughputs between samples, trimmed mean of M-values (TMM) algorithm (Robinson and Oshlack, 2010) from edgeR (R package) (Robinson et al., 2010) were used and CPM (count-per-million) was used as gene expression unit. Genes with CPM < 1 and not detected in at least 2 samples were excluded. Next, $\log_2$ transformation and quantile normalization were applied before data analysis.

Mutational Signature Analysis

Mutational Signature and Carcinogen Signatures

Based on the single nucleotide substitution and its' adjacent bases pattern of samples, frequencies of 96 possible mutation types for each sample could be estimated. Non-negative matrix factorization (NMF) algorithm was used to estimate the minimal components that could explain maximum variance among samples. Then each component was compared to mutation patterns of 30 validated cancer signatures reported from the COSMIC database and of 54 carcinogen signatures reported from Kucab et al. (Kucab et al., 2019) individually to identify cancer-related mutational signatures and carcinogen signatures. Cosine similarity analysis (Alexandrov et al., 2013b; Mayakonda et al., 2018) was used to measure the similarity between component and signatures, which ranged from 0 to 1, indicating maximal dissimilarity to maximal similarity. After decomposing matrix of samples' 96 substitution classes into 5 signatures, contribution of signatures in each sample could be estimated.

Estimation of APOBEC Enrichment Score

Deaminase effects of APOBEC were preferentially found in *TCW* motif, especially resulted in C to G or C to A substitution. Based on these characteristics, APOBEC enrichment score was calculated using the method reported by Roberts et al. (Roberts et al., 2013). Briefly, within 20bp of mutated bases, enrichment score for each sample is estimated by the overrepresented level of C>T and C>G substitutions within *TCW* motifs over all C>T and C<G substitutions and compared to the frequency of *TCW* in the background. Score estimation was also accounted reverse strand effects. Samples with enrichment score > 2 were grouped into the APOBEC enriched group.

Sample Preparation for Proteomic and Phosphoproteomic Analysis

Preparation of Reference Tissue Samples

Two reference tissue samples were prepared for quality control in each TMT batch experiment. The first reference sample was pooled from a total of 25 pairs of tumor tissue and adjacent normal tissues from early stage patients and label with TMT-126 (described below). Because this study focused on the early stage lung adenocarcinoma, the second reference sample was prepared by pooling 15 tumor tissues from late-stage patients and labels on of TMT-131 channel.

Protein Extraction and Tryptic Digestion

For TMT 10-plex proteomic and phosphoproteomic experiments, fresh frozen paired tumor and adjacent normal tissues (< 50 mg for each) were sliced into small pieces and washed by ice-cold PBS at least three times to remove blood. Sliced tissue specimens were re-suspended in 10-fold volume (10 mL for 1 mg tissue) of lysis buffer containing 100 mM Tris-HCl (pH 9.0), 12 mM sodium deoxycholate, 12 mM sodium lauryl sulfate, EDTA-free protease cocktail inhibitors, and phosphatase cocktail inhibitors and homogenized by Precellys 24 homogenizer (Bertin Technologies). The homogenized samples were heated at 95°C with vortexing at 750 rpm for 5 min and sonicated for 10 min (30 s on, 30 s off) using Bioruptor Plus (Diagenode, Denville, NJ). Residual debris was removed by centrifugation (16,200 x g for 30 min at 4°C). Then, methanol/chloroform protein precipitation was performed on the supernatant as follows: 1 volume of sample, 4 volumes of methanol, followed by an equal volume of chloroform were sequentially added with thorough mixing upon each addition. Then, three volumes of ultrapure water were added with intensive vortexing. After centrifugation (10 min at room temperature, at 16,200 x g), the upper aqueous phase was removed without disturbing the interface and 3 volumes of methanol were added with thorough vortexing. Followed by centrifugation at 16,200 x g for 10 min at room temperature, the supernatant was removed, and the white protein precipitate was allowed to air dry. The protein precipitate was resuspended in digestion buffer containing 8 M urea, 50 mM TEABC, EDTA-free protease inhibitors and phosphatase inhibitors. The protein concentration was determined using BCA Protein Assay following the manufacturer's protocol (Thermo Scientific).

For tryptic digestion, 450 μg proteins from individual sample spiked with an internal standard protein (yeast ALD4 protein, Sigma) at a ratio of 1:250 (w:w) was reduced by 5 mM dithiothreitol at 29°C for 30 min and alkylated with 10 mM iodoacetamide at 29°C for 45 min in the dark. The samples were diluted with 1-fold volume of 50 mM TEABC and reacted with Lys-C at a ratio of 1:100 (w:w, Lys-C:protein) at 29°C for 3 h. The samples were further diluted with 3-fold volume of 50 mM TEABC for trypsin digestion with a ratio of 1:50 (w:w, trypsin:protein) for 18 h at 29°C. The proteolytic digestion was stopped by adding 10% TFA to a final concentration of 0.5% in the sample. The digested peptides were desalted by SDB-XC StageTip protocol described previously (Dimayacyac-Esleta et al., 2015) followed by BCA assay for measuring peptide concentration.

TMT 10-plex Labeling

Desalted peptides from each sample were labeled with TMT10plex™ isobaric label reagents according to the manufacturer's instructions (Thermo Scientific) with a ratio of 0.8 mg reagent to 100 μg peptides. Synthetic β-casein standard phosphopeptide (FQpSEEQQQTEDELQDK) was spiked into tryptic peptides with a ratio of 1:250,000 (w:w) before TMT 10-plex labeling. For each set of TMT 10-plex, 126 and 131 channels were used to label the reference samples as described above. Four pairs of tumor and adjacent normal tissues were labeled with the other eight channels (adjacent normal tissues labeled with 127N, 128N, 129N, and 130N; tumor tissues labeled with 127C, 128C, 129C, and 130C). Two units of each TMT channel were freshly dissolved in anhydrous acetonitrile (ACN) with a ratio of 0.8:41 (w:v, mg:μL); 82 μL of each TMT channel in ACN was further added into 200 μg peptides dissolved in 200 μL of 100 mM TEABC. After incubation for 1 h at room temperature, the reaction was quenched by adding 16 μL of 5% hydroxylamine and incubated for 15 min. 2.5% of TMT-labeled peptides from each channel was taken out and subjected to LC-MS/MS analysis for determining labeling efficiency before pooling. Pooled 10 channels of TMT-labeled peptides were desiccated by Speed-Vac for High-pH RPLC fractionation.

Peptide Fractionation by High-pH RPLC

To increase the profiling depth of proteome and phosphoproteome, peptide fractionation was performed by high-pH reverse phase liquid chromatography (RPLC). Pooled and dried TMT 10-plex labeled peptides were re-dissolved in 600 μL of 5 mM ammonium formate (pH 10) and 2% ACN, and loaded on a 4.6 mm x 250 mm Zorbax 300 Å Extend-C18 column (Agilent, 3.5 μm bead size) at a flow rate of 0.5 mL/min on Waters alliance e2695 LC instrument coupled with 2489 UV/Visible detector and fraction collector III. Solvent A (2% ACN, 5 mM ammonium formate, pH 10) and a nonlinear increasing percentage of solvent B (90% ACN, 5 mM ammonium formate, pH 10) were used for peptide separation. A 120-min LC gradient run started with 100% solvent A for 10 min, then increased linearly to 12% B in 4 min, 40% B in 63 min, 60% B in 7 min, 90% B in 15 min, and maintained at 90% B for 8 min. Peptides were separated and collected every minute for a total of 96 fractions from 11 to 106 min and combined into 24 fractions with a stepwise concentration strategy. Following desalting with SDB-XC StageTip, 10% of each fraction were aliquoted and dried by vacuum centrifugation for proteome analysis. The remaining 90% of concatenated fractions were further combined into 12 fractions, dried with Speed-Vac and stored at −80°C until phosphopeptide enrichment.

Phosphopeptide Enrichment

The phosphopeptide enrichment was performed by home-made immobilized metal ion affinity chromatography (IMAC) StageTip capped at one end with a 20-μm polypropylene frits disk (Agilent, Wilmington, DE, USA) enclosed in the narrow end of tip fitting. For preparation of IMAC StageTip, 10 mg of Ni-NTA silica resin (QIAGEN) was dissolved in 225 μL of 6% acetic acid and loaded onto StageTip by centrifugation (3,300 x g for 3 min at RT). The Ni^{2+} ions were removed with 200 μL of 50 mM EDTA using centrifugation (300 x g for 4 min). The StageTip was then activated with 200 μL of 100 mM $FeCl_3$ and equilibrated with 200 μL of loading buffer (6% acetic acid, pH 3.0) prior to sample loading. Fractionated peptides were reconstituted in 0.5% acetic acid (pH 3.0) and loaded onto the IMAC StageTip with centrifugation (300 x g, 2 min). After successive washes with washing buffer A (80% ACN, 1% TFA) and washing buffer B (0.5% acetic acid), the bound peptides were eluted twice with 80 μL 200 mM $NH_4H_2PO_4$ and desalted using RP StageTips.

LC-MS/MS Analysis

Proteome Analysis

For proteome analysis, LC-MS/MS analysis was performed with a Thermo Scientific UltiMate 3000 RSLCnano system (Thermo Fisher Scientific) coupled to a Q Exactive HF (Thermo Fisher Scientific). The fractionated peptides were separated using Thermo Scientific PepMap C18 50 cm x 75 μm ID column (Thermo Fisher Scientific) with a 5%–28% ACN gradient in 0.1% FA over 200 min at a flow rate of 250 nL/min. The spectra of full MS scan (*m/z* 375-1600) were acquired in the Orbitrap mass analyzer at 120,000 resolution for a maximum injection time of 50 ms with an AGC target value of 3e6. Up to 15 precursors were selected for MS2 analysis with an isolation window of 0.7 Th and dynamic exclusion time was set to 20 s. Precursors were fragmented by HCD using a normalized collision energy of 33% and analyzed using the Orbitrap at 60,000 resolution for a maximum injection time of 120 ms with AGC target value of 1e5. Precursor ions with unassigned charge state as well as charge state of 1+, or superior to 7+ were excluded from fragmentation selection.

Phosphoproteome Analysis

For phosphoproteome analysis, LC-MS/MS analysis was performed with a Thermo Scientific UltiMate 3000 RSLCnano system (Thermo Fisher Scientific) coupled to an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific). Enriched phosphopeptides were dissolved in 1% formic acid and separated using Thermo Scientific PepMap C18 50 cm x 75 μm ID column (Thermo Fisher Scientific) with a 5%–28% ACN gradient in 0.1% FA over 200 min at a flow rate of 250 nL/min. The full MS spectra (*m/z* 375–1,600) were acquired in Orbitrap at 120,000 resolution with an AGC target value of 4e5 charges for a maximum injection time of 50 ms. Precursors for MS2 analysis were selected using a Top10 method with an isolation window of 0.7 Th. MS2 precursors were fragmented by HCD using a normalized collision energy of 36%. The AGC target value was set to 5e4 and the dynamic exclusion was set to 20 s. MS2 spectra were acquired in Orbitrap with a maximum injection time of 150 ms at a resolution of 60,000. Precursor ions with the charge state of unassigned, 1+, or superior to 7+ were excluded from fragmentation selection.

Mass Spectrometry Data Analysis

Database Search

All MS raw files from the same batch were processed together with Proteome Discoverer (ver. 2.1) by Mascot (ver. 2.3.2) and SequestHT engines (Eng et al., 1994) against the SwissProt human protein database (version 2016.05, 20,213 entries), combined with the spiked internal standard ALD4 protein from yeast (SwissProt: P46367) for proteome analysis and with β-casein protein (SwissProt: P02666) for phosphoproteome analysis. For phosphoproteome analysis, MS Amanda (Dorfer et al., 2014) was also included in Proteome Discoverer. The tolerance for spectra search allowed 10 ppm mass tolerance for precursor, 0.1 Da for product ions, and trypsin enzyme specificity with up to 2 missed cleavages. All search engines considered static carbamidomethylation (+57.022 Da) on Cys residues and TMT modifications (+229.163 Da) on the peptide N terminus and Lys residues, as well as dynamic oxidation (+15.995 Da) on Met residues, deamidation on Asn and Gln residues (+0.984Da), and acetylation on protein N terminus (+42.016 Da), while the phosphorylation on Ser, Thr and Tyr residues (+79.966Da) was additionally included as dynamic modification for phosphoproteome analysis. The minimal peptide length was set as 7 residues. The false discovery rate (FDR) of peptide and protein were all set as 1%. For phosphosite localization, the ptmRS (Taus et al., 2011) was used to determine phosphosite confidence and the phosphosite probability > 0.75 is considered as confident phosphosites.

Quantification of Global Proteome Data

The quantification analysis was performed at the protein level by Proteome Discoverer. The protein abundance was calculated from the sum of peptide abundance with quan value corrections and co-isolation threshold filtering < 50%. The protein-level TMT signal-to-noise (S/N) values of each batch were exported from Proteome Discoverer 2.1 and unique protein groups were consolidated in a single table containing all samples (columns = samples, rows = proteins). Protein S/N values were normalized by column-median to correct for equal loading across samples. The normalized S/N values were $\log_2$-transformed and the row-mean was subtracted from each protein for each batch separately to obtain $\log_2$-scaled values. The $\log_2$-scaled values were further transformed to column z-scores for Principal Component Analysis (PCA) in RStudio. Additionally, $\log_2$T/N values were computed using the normalized S/N values, which were further transformed to column z-scores and centered at zero (row-wise) to obtain the centered-$\log_2$T/N values. All normalization steps were performed in Perseus 1.6 (Tyanova et al., 2016).

Quantification of Phosphoproteome Data

The quantification analysis for phosphoproteome data was performed at phosphopeptide level by Proteome Discoverer. The TMT S/N values of phosphopeptides mapping to the same site were aggregated by median values. Phosphosite S/N values were normalized by column-median to correct for equal loading across samples and $\log_2$T/N values were computed. These were further transformed to column z-scores and centered at zero to obtain the centered-$\log_2$T/N values. All normalization steps were performed in Perseus 1.6 (Tyanova et al., 2016). To analyze the phosphorylation changes, the $\log_2$T/N value of each phosphosite was corrected for the respective protein $\log_2$T/N value using linear regression to obtain net phosphorylation changes.

Quality Control Evaluation

To evaluate the data quality of every batch of proteomics and phosphoproteomics datasets, we used proteins (n = 7,673) or phosphopeptides (n = 3,505) commonly identified in all proteomics or phosphoproteomics experiments for pairwise Pearson correlation

coefficient analysis. The log_2(TMT126/TMT131) ratio derived from column z-score transformed intensity were used to generate pairwise scatterplots for Pearson correlation and histogram of each proteomic and phosphoproteomic analysis by using psych R package.

Identification of Variant Peptides

Based on the detected somatic mutations in each patient and UCSC gene annotations, somatic mutations of each patient were annotated for their genomic locations and their impact on protein functions was predicted by customProDB (R package) (Wang and Zhang, 2013). Accordingly, we retrieved the nonsynonymous mutations in each patient and generated patient-specific mutated protein sequences. For variant peptide identification, we combined the mutation protein sequences from patients in the same batch as a combined customized mutation database.

All MS raw files were processed using Proteome Discoverer (ver. 2.1) with SequestHT search engines against the SwissProt human protein database (version 2016.05, 20,213 entries), combined with batch-specific mutated protein database derived from the WES data from 4 patients in the same batch. The tolerance of all spectra allowed ± 10 ppm mass tolerance for precursor, +/− 0.1 Da for product, and up to 2 missed cleavages for trypsin enzyme specificity. All search engines considered static carbamidomethylation (+57.022 Da) on Cys residues and TMT modifications (+229.163 Da) on the peptide N terminus and Lys residues, as well as dynamic oxidation (+15.995 Da) on Met residues, deamidation on Asn and Gln residues (+0.984 Da), and acetylation on protein N terminus (+42.016 Da). Peptide identification stringency was set at a maximum FDR of 1% and a minimum peptide length as 7 residues. Peptide spectral matches (PSMs) were validated using percolator based on q-values at 1%. Unique proteins with < 1% FDR were considered as positive identifications. The PSMs of identified variant peptides were manually confirmed. We also cross-checked the TMT reporter intensity in the variant PSM and the patient-derived variant database to identify the variant-carry patient. The identified variant peptides that had isobaric substitutions were excluded.

Multi-omic Data Analysis

Identification of pQTL and eQTL

We tested for associations between the RNA and protein relative abundances against a set of 88 variant genes, filtered for at least 5 events across the adenocarcinoma patients using eQTL and pQTL analysis. Patients with hypermutation profiles and proteins with missing values were excluded. Genes on X, Y and mitochondrial chromosome were also excluded. We finally used 27,768 RNAs and 6,997 proteins for the QTL mapping. Both datasets were quantile normalized to a Gaussian distribution before fitting a model. All associations were performed by LIMIX (Chen et al., 2016) using a linear regression test. Let the quantitative measurements be y, xs corresponds to somatic variants in binary format. The linear regression model is as follows:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{x}s\beta s + \boldsymbol{\psi}, \text{ where } \boldsymbol{\psi} - (0, \sigma \boldsymbol{I}).$$

Here βs denotes the effect size of the tested variant, μ is the intercept and ψ is the residual noise.

Kinase Activity Prediction

Inference and prediction of putative kinase activity associated with high APOBEC mutation signature is performed from the phosphoproteomic data using PHOXTRACK (Weidner et al., 2014). In brief, the phosphopeptide entries quantified in more than 50% of patients were used and their mean of log_2(APOBEC-high/APOBEC-low) was subjected into PHOXTRACK analysis with the criteria of 1000 permutations, at least 5 phosphosites per kinase, and unweighted statistic *p-value* < 0.05.

Consensus Clustering

Unsupervised clustering of the patient samples at the different molecular levels was performed with the ConsensusClusterPlus R package (Wilkerson and Hayes, 2010) using the top 50% most variable proteins, RNAs and phosphosites based on standard deviation ranking. For the proteomics and RNA data, genes with missing values were pre-excluded whereas for the phosphoproteomics data with less than 20% missing values, we imputed the missing data from the normal distribution within a range of 20% of the standard deviation of all values in Perseus 1.6. Patient clusters were derived based on k-means clustering, Euclidean distance and 1,000 resampling repetitions in the range of 2 to 7 clusters. The empirical cumulative distribution function (CDF) plot initially showed optimal separation between 3 and 4 clusters for all the molecular levels. To assess the stability of the k = 3 and k = 4 consensus clusters we made silhouette plots in Rstudio using the ''cluster'' library. In all cases, the clusters at k = 4 were overall more stable except the 4th RNA cluster which was excluded. The 4th proteome cluster was also excluded due to its small size.

Weighted Correlation Network Analysis

Weighted correlation network analysis was performed with the WGCNA library (Langfelder and Horvath, 2008) in RStudio using the centered-log_2T/N values of 9,072 proteins quantified in at least 80% of the adenocarcinoma patients. A soft threshold at power 5 was selected based on scale free topology model fit (R^2 = 0.8). Module identification was performed with the ''cutreeDynamic'' function using the ''tree'' method and minModuleSize = 5. Functional annotation of the modules was performed with Fisher's exact test using GOBP-slim, GOCC-slim, CORUM and KEGG terms in Perseus 1.6 (Tyanova et al., 2016). Protein networks were visualized in Cytoscape (Shannon et al., 2003). For concise representation we trimmed the network keeping only the proteins that belonged to the top enriched terms from each annotation category, edges with correlation weight greater than 0.03 (positive associations only) and allowing less than top 50 interactions per source and target node.

HLA Genotyping Analysis and Neoantigen Prediction

Alleles with 4-digit HLA class I genes were inferred using POLYSOLVER (https://software.broadinstitute.org/cancer/cga/polysolver) (Shukla et al., 2015). Based on the somatic mutation information, mutated DNA sequences were translated into mutated peptide sequences. Mutated peptide sequences with length 8 to 14 amino acids were further input to NetMHC (v4.0) (http://www.cbs.dtu.dk/services/NetMHC/) (Andreatta and Nielsen, 2016) to predict corresponding binding affinities with MHC class I genes. The predicted binding affinity IC_{50} < 50 nM was denoted as strong binding and IC_{50} between 50-500 nM was weak binding.

Immune Cell Profiling

Immune cell profiling was performed in CIBERSORT (Newman et al., 2015) using the read count RNA-seq data for adenocarcinoma tumor and normal adjacent samples against the LM22 signature gene file encompassing 22 immune cell types. The analysis was done with 100 permutations in relative and absolute modes using internal quantile normalization of the data.

Stromal Scores

Stromal scores were computed with the ESTIMATE package (Yoshihara et al., 2013) in RStudio using quantile normalized RNA read counts for tumor and adjacent normal tissues.

Survival Analysis

We performed the survival analysis for APOBEC signature in TCGA Cohort Genomic datasets and clinical phenotype of TCGA lung adenocarcinoma were downloaded from the University of California Santa Cruz genome browser (https://xenabrowser.net/). APOBEC enrichment scores were estimated based on somatic SNVs. After samples were separated into APOBEC-high or -low group, overall survival (OS) analysis was performed to compare the two groups for evaluating prognosis of the APOBEC signature.

We also compared the disease-free survival (DFS) and overall survival (OS) among different EGFR genotypes in complete surgically resected stage IA and IB lung adenocarcinoma patients or MMPs IHC staining. Univariate analyses of patients' characteristics were performed using the Fisher's exact test and the independent t test. The Kaplan–Meier method was used to estimate DFS and OS. Differences in survival time were analyzed by the log-rank test. The Cox proportional hazard model was used to evaluate the impact of adjuvant chemotherapy for multivariate analyses of survival time. All statistical tests were carried out using SPSS 15.0 (SPSS Inc., Chicago, IL, USA). Two-tailed tests and p values < 0.05 for significance were used.

Immunohistochemical Analysis

To detect the expression of MMP2, MMP7 and MMP11 protein in the tissue by immunohistochemical (IHC) staining, 4-μm-thick sections from each formalin-fixed, paraffin-embedded (FFPE) tissue block were de-waxed with xylene and rehydrated through a graded series of ethanol. Antibodies for MMP2 (Clone 4D3, Santa Cruz Biotechnology; dilution 1:50 for 1.5 h), MMP7 (Clone JL07, Santa Cruz Biotechnology; dilution 1:100 for 1.5 h) and MMP11 (clone SL3.05, Santa Cruz Biotechnology; dilution 1:150 for 1.5 h) were used for immunohistochemistry. Antigen was retrieved by autoclaving for 10 min at 121°C in Novocastra Epitope Retrieval Solution pH 6 (Leica Biosystems). The UltraVision Quanto Detection System HRP DAB (Thermo Fisher Scientific) was used according to the manufacturer's instructions. Finally, the sections were counterstained with hematoxylin and then mounted.

QUANTIFICATION AND STATISTICAL ANALYSIS

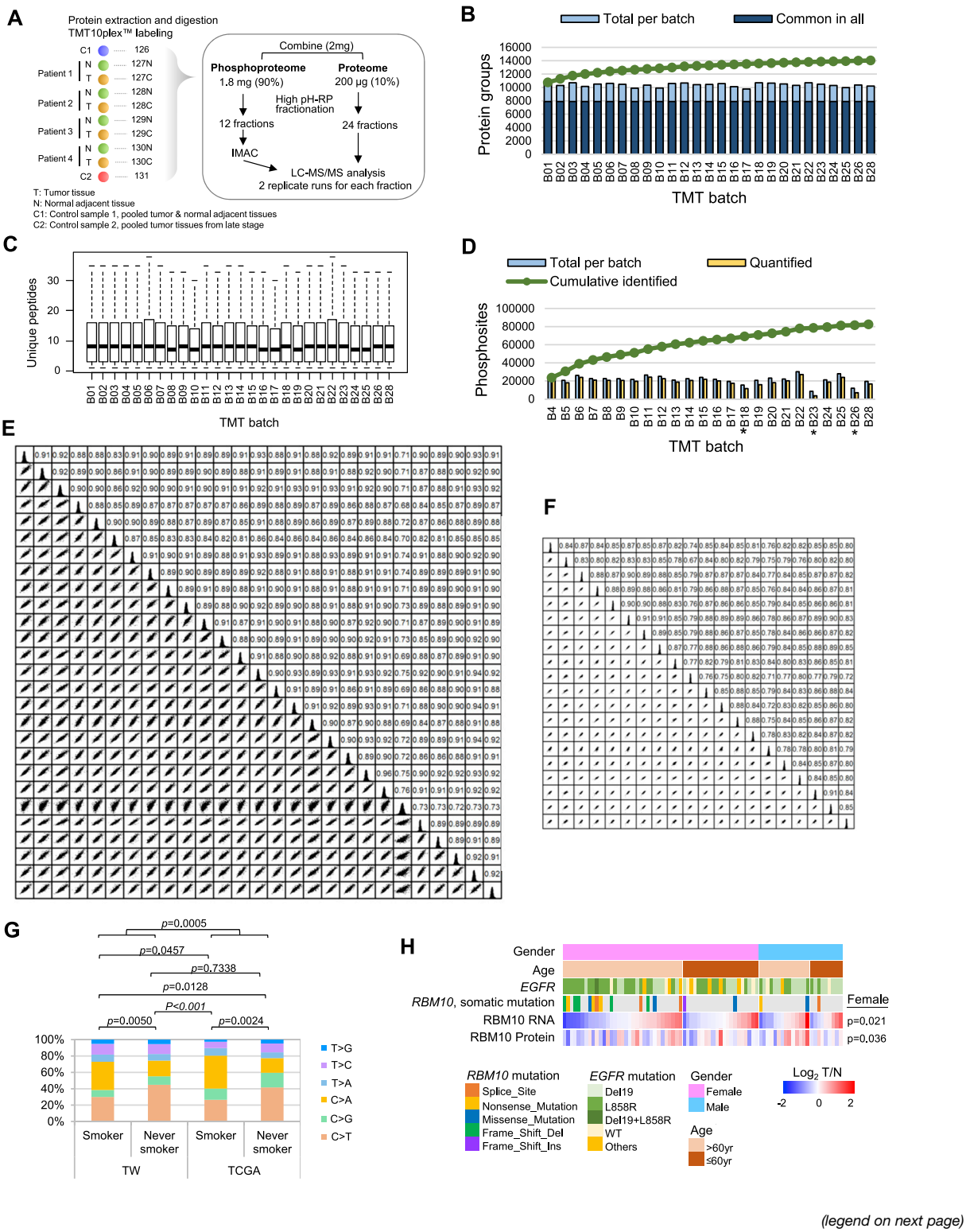
Methods of quantification and statistical analysis for all experiments and omic analyses were described in the Results, figure legends, and Method Details subsections.

Additionally, biological enrichment for molecules and pathways from multi-omics data was performed in Perseus 1.6 software (Tyanova et al., 2016) using the 1D-annotation pathway enrichment method (Cox and Mann, 2012) or with Fisher's exact test. The enrichment score indicates whether the proteins in a given biological term tend to be systematically up- or downregulated based on Mann-Whitney U test. The 1D-annotation enrichment method was also applied for the enrichment of KEGG pathways with low and high mRNA-to-protein correlations. The mean log_2T/N values and FDR on the expression of mRNA, proteins, and phosphopeptides of individual patients was computed by 1D-annotation pathway enrichment analysis. All enriched pathways were filtered for Benjamin-Hochberg FDR < 0.05. Next, the patients were annotated with 6 carcinogen signatures, including PAHs, nitro-PAHs, Nitrosamine, mixed, alkylating agents and radiation. The mean expression value within each enriched pathway from individual patients was further compared among different carcinogen groups using Kruskal-Wallis rank test and selected with p < 0.05. The enriched pathways had significantly higher expression in at least one carcinogen cohort.

For the assessment of the impact of mutations on protein and mRNA abundances, analysis of variance (ANOVA) was performed in Perseus 1.6 using preformatted tables. Permutation-based FDR correction was applied to adjust p values from the ANOVA. The web-based tool Morpheus (https://software.broadinstitute.org/morpheus/) was used for hierarchical clustering and visualization of heatmaps. For significant protein enrichment, Student's t test or Welch's t test, as the unequal variances of testing groups, was performed to test the difference in continuous variables. All tests were performed by two-tailed test and *p-value*s < 0.05 were considered significant.

CellPress
OPEN ACCESS

Supplemental Figures



(legend on next page)

Figure S1. Proteogenomics Workflow, Quality Control, and Comparison of Somatic Mutation Profiles between TW and TCGA Cohort, Related to Figure 1

(A) Overview of proteomics experimental design. (B) Bar plots showing the number of identified proteins per TMT batch. (C) Boxplots showing the median number of unique peptides identified per TMT batch. (D) Bar plots showing the number of identified and quantified phosphosites per TMT batch. TMT batches with low phosphoproteome coverage are highlighted with a mark and were excluded from downstream analysis. (E) Pearson's correlation coefficient between the ratio of two reference channels (126 versus 131) from each two different batches evaluated with 7673 proteins without missing values are shown in the upper triangular matrix. Distributions of fold changes derived from two different batches for these proteins are shown in the lower triangular matrix. Plots along the diagonal shows the distribution of $\log_2$ fold changes of these proteins of a single batch. (F) Pearson's correlation coefficient from each two batch of phosphoproteomics data using 3505 phosphopeptides without missing values. (G) Comparison of the proportion of DNA substitution mutations between TW and TCGA LUAD cohort. Fisher's exact test was performed for statistical analysis. (H) Clustergrams of gender-specific *RBM10* mutation profile. The somatic mutation, mRNA and protein expression levels (relative $\log_2$T/N ratio) of RBM10 were clustered with gender, age, and *EGFR* status. Fisher's exact test was performed for statistical analysis.

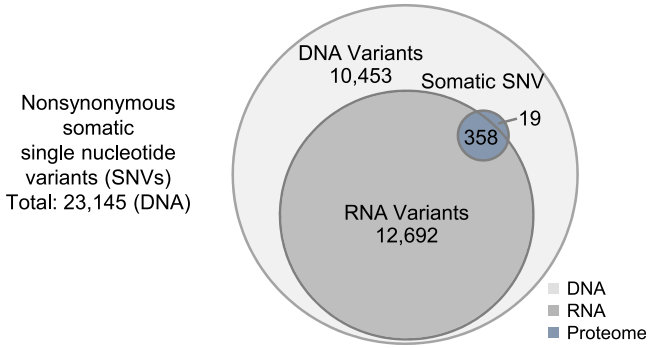


Figure S2. Summary of Sequence Variants Identified in DNA, RNA, and Protein Level, Related to Figure 2

Venn diagram showing the overlap of sequence variants identified at the DNA, RNA and protein levels.

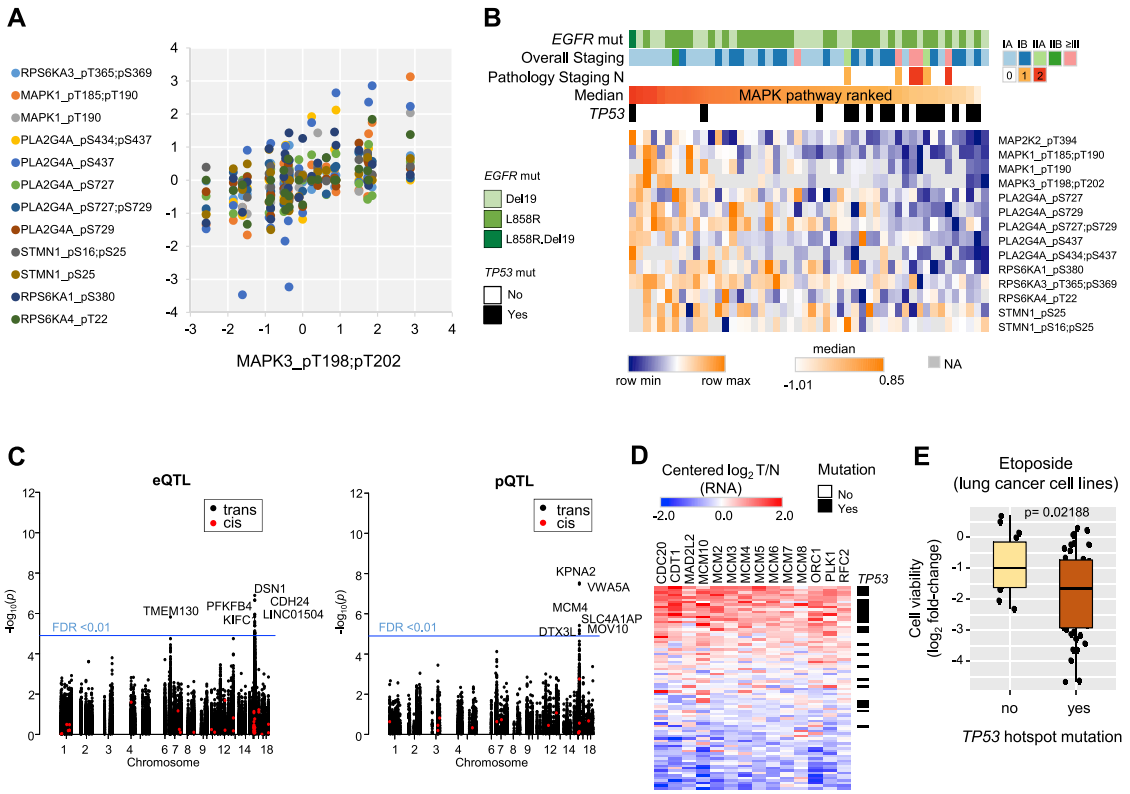
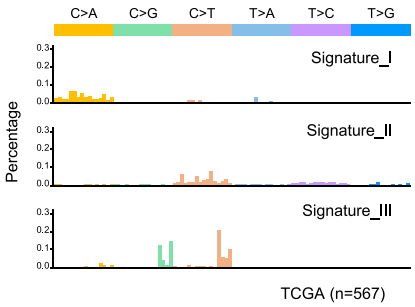


Figure S3. Impact of Genomic Alterations in the Proteome and Phosphoproteome of Lung Adenocarcinoma, Related to Figure 3

(A) Correlation between phospho-MAPK3 and its downstream phosphoproteins. (B) Ranked co-phosphorylation signature of the MAPK cascade of nonsmokers with different *EGFR* mutation status. (C) Manhattan plots of the most confident non-redundant association tests in eQTL and pQTL analysis. (D) Heatmap of cell cycle related genes associated with *TP53* mutations at the RNA level. (E) Drug response of lung cancer cell lines with *TP53* mutation from TCGA and COSMIC hotspot (Wilcoxon rank-sum test, p = 0.021).



	Reference mutational signature[1, 2]	Cosine similarity	Description
I	Signature_4	0.976	exposure to tobacco (smoking) mutagens
	Signature_26	0.932	DBA (75 uM) + S9 [PAHs]
II	Signature_6	0.873	defective DNA mismatch repair
	Signature_6	0.567	DMH (11.6 mM) + S9 [Alkylating Agents]
III	Signature_2	0.850	APOBEC Cytidine Deaminase (C>T)
	Signature_33	0.471	SSR (1.25 J) [Radiation]

[1]COSMIC [2]Kucab JE, et al. 2019

Figure S4. Summary of APOBEC and Carcinogen Signature in TCGA LUAD Cohort, Related to Figure 4

Trinucleotide motif frequency plots and enriched mutational signatures for three mutational profiles identified in the TCGA LUAD cohort.

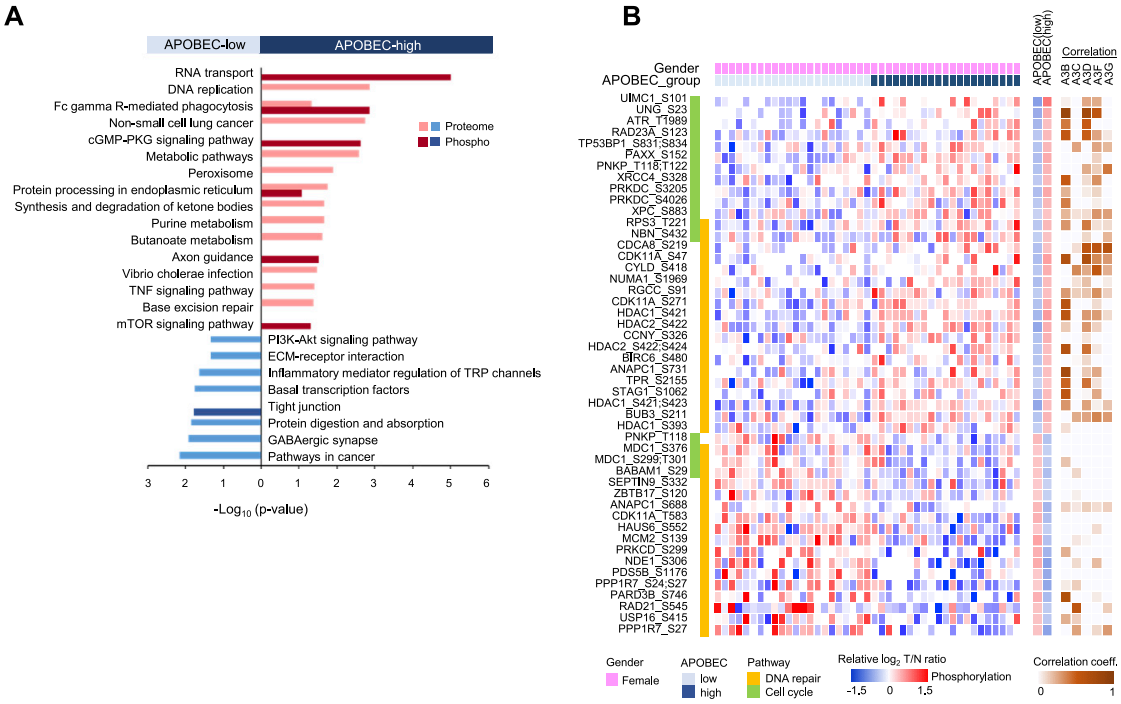


Figure S5. APOBEC Signature-Associated Pathways, Related to Figure 5

(A) Pathways enriched with altered mean values between APOBEC-high and -low groups in proteome and phosphoproteome level (Fisher's exact test, $p < 0.05$).

(B) Heatmap of phosphorylation sites with differential regulation associated with cell cycle and DNA damage enriched in APOBEC signature (t-test, $p < 0.05$). Correlation of log_2T/N ratio of these phosphosites with APOBEC family was analyzed by spearman correlation.

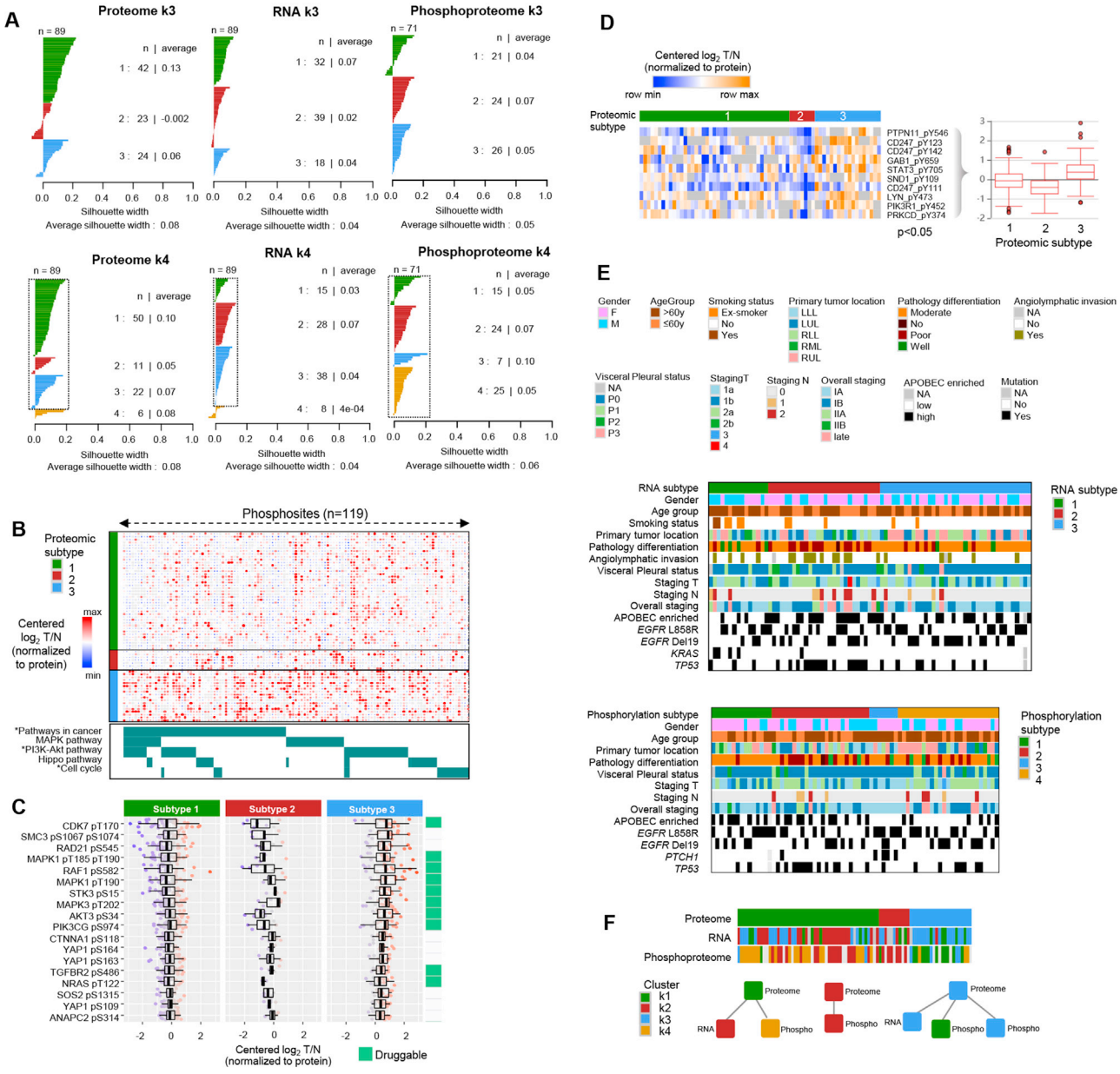


Figure S6. Multi-omic Subtypes, Related to Figure 6

(A) Silhouette plots for the different clusters identified by consensus clustering for proteomics, RNaseq and phosphoproteomics. The clusters selected for downstream analysis are highlighted with a dashed outline. (B) Heatmap of differentially regulated phosphosites (columns) between the three proteomic subtypes using protein-corrected phosphorylation abundances (ANOVA FDR < 0.1). Rows represent patients and larger dot size indicates higher relative phosphorylation. Only phosphosites from proteins in cancer related KEGG pathways are shown. The respective pathways are shown in the bottom panel. The asterisk indicates significant enrichment over the non-differentially regulated phosphosites (Fisher's exact test, Benj. Hoch. FDR < 0.1). (C) Boxplots of representative phosphosites with upregulation in the early stage subtype 3. The phospho-proteins that can be targeted by known inhibitor (source: DGIdb, http://www.dgidb.org/) are highlighted with a green mark. (D) Heatmap and boxplot of representative phospho-tyrosine sites with upregulation in the early stage subtype 3. (E) Alignment of RNA and phosphoproteomics subtypes with clinical features. F) Overlap between proteomics, RNA and phosphoproteomics subtypes. Significantly over-represented RNA and phosphorylation subtypes in each one of the proteomics subtypes are shown as networks (Fisher's exact test, p < 0.05).

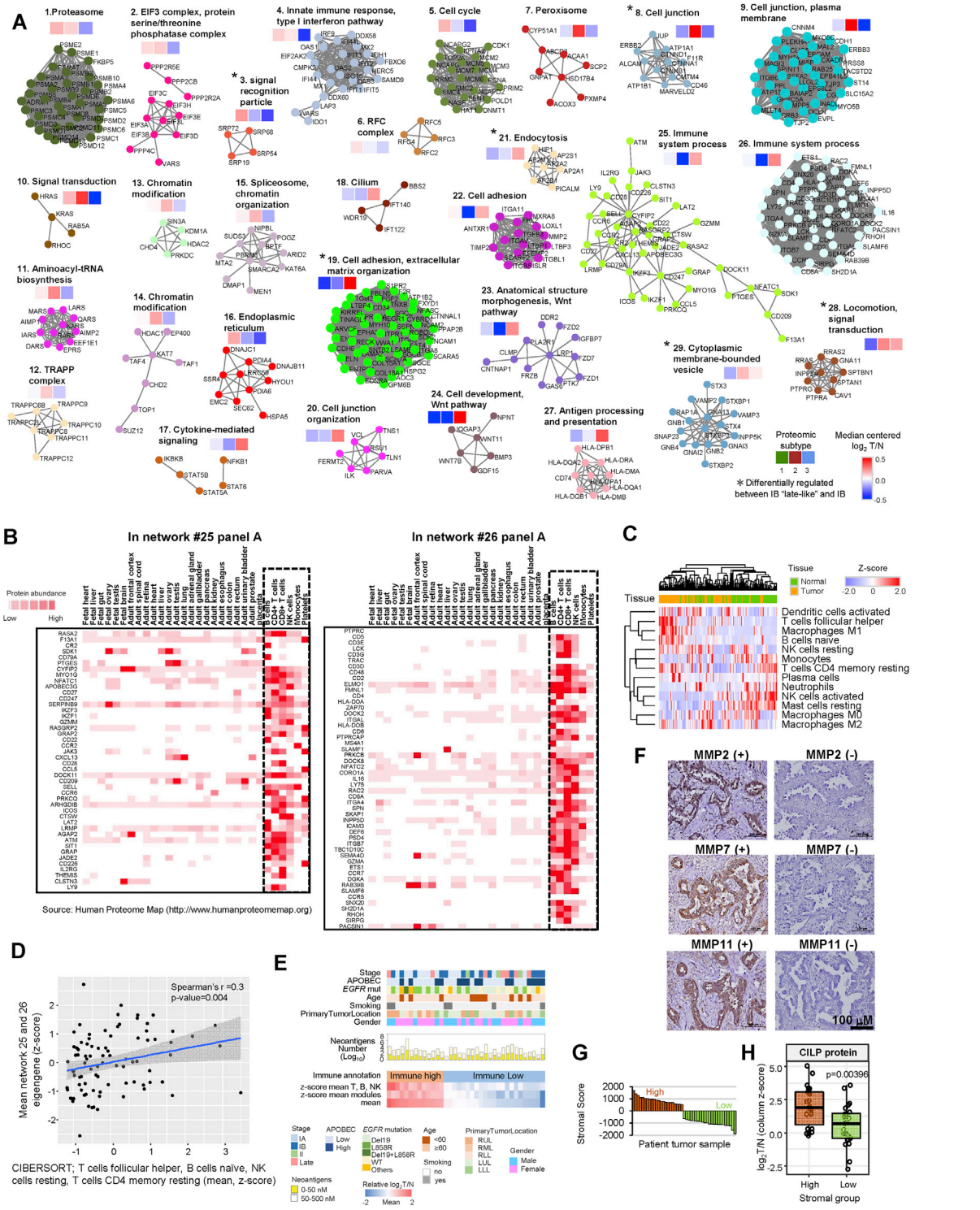


Figure S7. Differentially Regulated Networks and Immune Profiles between the Proteomic Subtypes, Related to Figure 7

(A) Protein networks with differential regulation between the proteomic subtypes based on eigengene values (ANOVA, FDR < 0.05). The edges in the networks represent *de novo* protein correlations and therefore not all nodes correspond to the biological terms shown on top. Protein networks with significant differential regulation between the IB and IB "Late-like" refined classes are indicated with an asterisk. (B) Tissue protein expression for the proteins found in networks #25 and #26 based on the Human Proteome Map database (https://www.humanproteomemap.org/). (C) Immune cell profiling based on CIBERSORT analysis based on the RNA-seq data. The heatmap shows the z-score transformed absolute immune cell abundances across patient tissues. (D) Correlation plot of the mean

(legend continued on next page)

z-score CIBERSORT results for T, B, NK cells versus the mean eigengene values of immune related networks #25 and #26. (E) Clustergram of immune cells-related protein profile. Clinical information was clustered by immune high and low cohort. Significant clinical features were further calculated by Fisher's exact test. (F) Immunohistochemistry staining for selected matrix metalloproteinases in an independent retrospective Taiwan lung adenocarcinoma cohort. (G) Stromal scores for classifying patients into two groups with high- and low- stromal scores (< 1st and > 3rd quartile) using quantile normalized RNA read counts for tumor and normal adjacent tissues. (H) Differential $\log_2$T/N values of CILP protein in the high- and low-stroma group (fold-change = 2.6, ANOVA, p = 0.00396).